

Performance Analysis of Models Used to Predict Failure in Secondary School*

Sofia Jordão¹, Dalila Durães¹[0000–0002–8313–7023], and Paulo Novais¹[0000–0002–3549–0754]

ALGORITMI Centre/LASI, University of Minho, Guimarães, Portugal
pg47489@alunos.uminho.pt, dad@di.uminho.pt, pjon@di.uminho.pt

Abstract. According to data made available by Pordata for the year 2021, about 8.3% of medium school students in Portuguese schools fail or drop out of the educational system, another 9.8% of students in this situation are still in basic education. Since education is one of the pillars of a country's development, it is important to understand the reasons behind these statistics and discover what leads students to such failure in order to try to mitigate these results. In order to do so, it is necessary to acquire data about the students, thus emerging the area of Educational Data Mining. Early prediction of school failure can be the key piece of the effort to avoid it. So, this paper presents a comparative study of machine learning models to indicate the best model to predict school failure.

Keywords: Artificial Intelligence · Machine Learning · School failure.

1 Introduction

Education is a process of acquiring knowledge, skills, values, and habits through various forms of learning. It is essential for human development and is one of the pillars for the integration of an individual into society. The dropout rate in education and training has been the main indicator, at the European level, of the evolution of educational systems, as its centrality is recognized in today's societies, for economic competitiveness, social cohesion, and equal opportunities [1].

If a student is at risk of failure, it is important to identify the underlying cause and treat it to improve their chances of success. The challenge is to develop technologies to predict the risk of student failure in advance in order to prevent it. Thus emerged the concept of Educational Data Mining (EDM), which focuses on the application of data mining techniques in the field of education. EDM is mainly used to develop accurate models that predict student performance in order to improve learning experiences, presenting great potential to improve the quality of educational institutions and systems [2].

The use of machine learning models in education has become increasingly prevalent in recent years and presents great potential to significantly improve

the decision-making process and increase student success. However, there is a constant debate about which models are most effective in the context of education. Given this focus, the main research question to be addressed is: Which machine learning models achieve better results in the field of education?

This paper discusses a study conducted to determine the most effective models for preventing failure in secondary education. The results of different models are compared to determine the best approach.

The paper is structured into six main sections. The second section provides an overview of the current state-of-the-art in preventing school failure. The dataset with the pre-processing and methodology are explain on section III. Section IV explain the Experiments applied. The results of the study are presented and discussed on section V, leading to a conclusion at section VI.

2 Background

Educational Decision Support Systems (EDSS) are computer-based systems that help educators and educational organizations make informed decisions about teaching and learning. An EDSS may use data mining techniques to analyze student performance data in order to identify patterns and trends that can be used to support decision making related to their learning. It may also use data visualization tools to present this data in an intuitive manner, allowing educators to make informed decisions.

Although EDSS has been widely studied in higher education, its capabilities for primary and secondary education have been little explored. In higher education, EDSS has been mostly used to support resource allocation and learning analytics, as well as to analyze student performance data to identify factors contributing to their success [3, 4].

Phauk Sokkhey and colleagues presented a new web-based system for predicting low-performing students using data mining techniques in their study [5]. A dataset with 43 features was created through a carefully designed questionnaire that covered personal information, domestic factors, individual factors, school factors, and score records. The authors used various filter-based selection methods, including a new method combining Chi-square and Mutual information, to create an optimal set of features for easier training. However, the authors did not disclose the features in this optimal set. Four different algorithms, including K-Nearest Neighbour, Hybrid C5.0, Hybrid RF, and an Improved Deep Belief Network optimization approach built by the authors, were tested to build the model. Hybrid RF was the best-performing algorithm for both the normal and optimal sets, achieving an accuracy of 99.95% and 99.98%, respectively. Additionally, the proposed filter-based selection method produced the best results.

J. Wang and colleagues [6] conducted a study to identify the key factors that contribute to academic success of students in the United States from their first year of school up to the end of high school. They analyzed a dataset that included information on school enrollment, student finances, and academic performance. Using a Random Forest-based data mining model, the authors achieved a mean

square error of 0.910, outperforming other algorithms like Lasso and ElasticNet. The authors determined that the number of students enrolled in a class and the location of the school were the most significant factors influencing students' grades.

S. Begum *et al.* [7] conducted a study to predict students' performance by implementing algorithms. Two datasets were created for this purpose: one for mathematics, with 395 examples, and another for Portuguese classes, with 649 entries. Three algorithms were developed along with an optimized version of each. The algorithms used were Boosting, Bagging, and Random subspace algorithms. Genetic Algorithms were developed to optimize each algorithm. Boosting was found to be the best algorithm for predicting success in both mathematics and Portuguese, with an accuracy of 98.11% and 96.82% respectively.

3 Dataset

Regarding the original dataset, it consisted of 123 features containing information ranging from more general information about the student, such as gender, age and grades, to more personal information such as their satisfaction with their grades and their level of school involvement. Table 1 present some details of the dadta from the dataset.

Before working with the dataset, it was necessary to pre-process it in order to ensure data quality.

Firstly, since this data comes from a questionnaire conducted by students under the age of 16, it was necessary to ensure that only data to which we were allowed to use would be used. After that selection we were left with 466 students data. These students encompass five academic years, with each class consisting of approximately 20 students. To elaborate, the 5th and 6th class have five classes each, totaling approximately 200 students, evenly distributed. As for the 7th, 8th, and 9th grades, there are a collective 266 students, divided into classes of roughly 20 students each.

The dataset underwent various data preparation steps following the CRISP-DM framework [8]. Initially, efforts were made to handle missing values and identify duplicate entries. Subsequently, the data needed to be transformed into numerical format to make it compatible for model input. Redundant data within the dataset was also removed.

After a first analysis of the dataset, we found that there were 28 columns with null values. The first ones consisted of how many days a week the students had tutoring and what type of tutoring they had. The reason why there were so many null values was because most students do not have tutoring and therefore left the question blank. The solution then was to fill in these null values with 0. There was also one student who did not answer what their grade was in Portuguese and Mathematics for the 1st period. In order to not discard the remaining information provided, these missing values were replaced by the average grade in both subjects. Furthermore, the remaining columns were mostly empty and did not provide significant information, so they were discarded.

Table 1. Dataset features example

Personal Data		Socio-biographical Data	School Data	Questionnaire
Name		School Allowance level	Name of the Course	"If it's important for students to show that they're smarter"
Gender		House Situation	Name of Subject	"Whether it's important to avoid showing that you're having difficulties with your work"
Age	Father's Educational Qualifications		Subject Grade	"If you hand in your homework on time"
Naturalty	Mother's Educational Qualifications		Average Grade	"Participates actively in class"
Town	Transport for School		Self-Evaluation	"Do you tutor? If yes, how many days a week do you go and what type of tutoring?"
School Year	Parents' marital status		School Situation	"How satisfied are you with your grades?"

In the survey, there were multiple questions related to private tutoring, including whether students received it and how frequently. This information was consolidated into a single column, where "1" denoted students without tutoring, "2" indicated those receiving tutoring once a week, and so forth. Columns like "username" and "class," primarily used for student identification, were dropped. Eliminating these columns not only removed irrelevant information for the model but also ensured the students' anonymity was preserved.

Upon examining the grade distribution, it was noted that the data was imbalanced. To address this, oversampling techniques, as described in [9], were applied using the RandomOverSampler from the imblearn.over_sampling library.

The resulting dataset was split into two subsets: one for predicting Portuguese grades and another for predicting Math grades. Both subsets contained the same demographic information, but the Portuguese dataset exclusively included final Portuguese grades as the target variable, while the Math dataset focused solely on students' final Math grades.

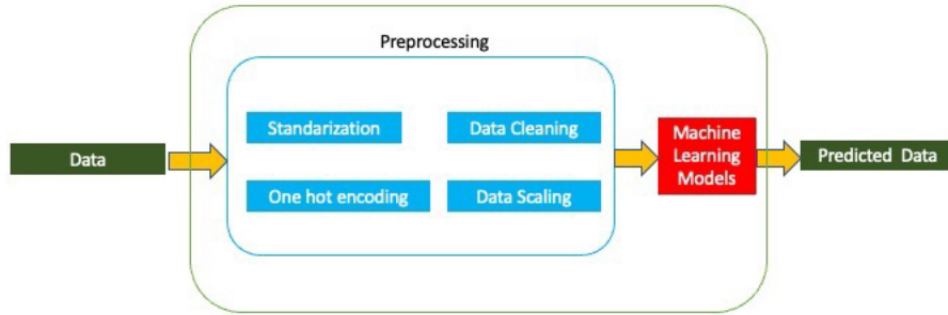
To tune the models' hyperparameters, GridSearch was used which improved its performance. So, for the 123 initial features, we obtain only 6 important features for obtain the level of student risks. The final resume of the final dataset is presented on Table 2.

Table 2. Final Dataset Features

Feature	Name
1	Results of Portuguese on 2nd Trimester
2	Results of Math on 2nd Trimester
3	satisfaction of the student about the results
5	Assessment of the students on the questionnaires
6	Student involvement in school

3.1 Methodology

This subsection explains the methodology for predict student failure. It includes the pre-processing steps and the model used for prediction. A visual representation of the entire system is provided in Figure 1. The system architecture comprises three main components, namely packet input, pre-processing, and ML model.

**Fig. 1.** Methodology for predicted failure.

To begin with, the dataset values were standardized to ensure that they were in the correct format. The data was cleaned by removing any corrupted values. The next step involved implementing one-hot encoding due to the presence of multiple values in the categorical features. This process resulted in the creation of N binary columns, where N represents the number of distinct values in a given column. A value of 1 in a binary column indicated the presence of that value in the current instance, while a value of 0 indicated its absence. Prior to this step, data cleaning was conducted to eliminate any corrupt or irrelevant values from the dataset. The four step in our pre-processing involved data scaling, which we achieved using the Min-Max scaler. Scaling is a technique that allows for the transformation of the values of numeric variables in a dataset into a common scale. Finally, it was applied the machine learning models with one with the predicted data.

4 Experiments

This section details the experiments that were carried out to confirm the effectiveness of the approach outlined in Figure 1. We provide an overview of the experimental setup, including the methodology used to validate our models. Additionally, we present the metrics used and illustrate the outcomes of the experiments.

4.1 Metrics

To make sure that our predictive models are effective, we assess their performance using F1-score, Precision, and Recall. Since our dataset is imbalanced, where accuracy alone could lead to inaccurate results, these metrics are crucial. F1-score, Precision, and Recall offer a more precise evaluation of the model's performance by addressing the imbalance in the dataset and guaranteeing a just evaluation.

4.2 Machine Learning Models

Once the dataset was preprocessed, it was time to use machine learning models to predict students at risk. Remembering that the target class is the cluster. Each student, based on the previously selected features, will be assigned a value from 0 to 4, where 0 and 1 represent students at high risk of school failure and 2 to 3 represent students at low risk. Taking into consideration the type of data to be worked with and the objective of this work, supervised learning models were used. The selected models were Hybrid Random Forest [10], Decision Tree [11], XGBoost, and Boosting [12], as they are the most commonly used in this domain according to the literature.

For the Hybrid Random Forest model it was used the Random Forest Classifier algorithm as well as the Principal Component Analysis also known as PCA, and cross validation [10]. As for Boosting algorithms it was used Gradient Boosting and AdaBoost [12].

In all the models, Grid Search was used in order to find the best parameters for each model. Cross-validation, with 10 folds, was also executed in all the models to ensure that there was no overfitting.

5 Results and Discussion

In this section, the results of each model for each cluster will be presented, displaying their confusion matrix and a classification report with the values of the metrics precision, recall, and F1-score.

The first model to be developed was the hybrid random forest. For this model the first step was to standardize the data, after that it was applied the PCA algorithm. After doing so the Random Forest Classifier algorithm was applied while using Grid Search for the best parameters. The second model to

be developed was the decision tree. The third model to be developed was the XGBoost Classifier. The fourth model to be developed was the Gradient Boosting Classifier and the fifth model to be developed was the ADA Boost Classifier.

Table 3. Classification report - Cluster 0

Model	Precision	Recall	F1-score	Support
Hybrid Random Forest	0.98	0.98	0.98	44
Decision Tree	0.91	0.89	0.90	44
XGBoost	0.98	1.00	0.99	44
Gradient Boosting	0.95	0.93	0.94	44
AdaBoosting	0.94	0.77	0.85	44
accuracy			0.93	44

Table 4. Classification report - Cluster 1

Model	Precision	Recall	F1-score	Support
Hybrid Random Forest	0.95	0.95	0.95	40
Decision Tree	0.86	0.95	0.90	40
XGBoost	1.00	0.97	0.99	40
Gradient Boosting	0.93	0.97	0.95	40
AdaBoosting	0.78	0.95	0.85	40
accuracy			0.96	169

When it comes to predicting the students with higher risk of school failure, students which belong in both 0 or 1 cluster, we can observe by the Tables 3 and 4 that XGBoost obtained overall better metrics. Also by observing all its results on the prediction of different clusters we can verify that this model achieved an overall accuracy of 99% before cross validation. And approximately 96,4% after cross validation.

Table 5. Classification report - Cluster 2

Cluster	Precision	Recall	F1-score	Support
Hybrid Random Forest	0.93	0.95	0.94	40
Decision Tree	1.00	0.88	0.93	40
XGBoost	0.98	1.00	0.99	40
Gradient Boosting	0.97	0.88	0.92	45
AdaBoosting	0.91	0.97	0.94	40
accuracy			0.92	169

Table 6. Classification report - Cluster 3

Cluster	Precision	Recall	F1-score	Support
Hybrid Random Forest	1.00	0.98	0.99	45
Decision Tree	0.94	0.98	0.96	45
XGBoost	1.00	0.98	0.99	45
Gradient Boosting	0.94	1.00	0.97	45
AdaBoosting	0.98	0.89	0.93	45
accuracy			0.99	169

As for predicting students with lower risk of school failure, belonging in cluster 2 or 3, we can observe that XGBoost still maintains overall better metrics than the other methods. For cluster 3, hybrid random forest model obtained equal metrics to XGBoost but when compared to the other clusters, we verify that this model only achieved an overall accuracy of 96% before cross validation. And approximately 97% after cross validation.

Regarding the other models, such as Decision Tree, Gradient Boosting and AdaBoost, they achieved an overall accuracy of 94,5%, 94,7% and 89,4% after cross validation, respectively.

5.1 Which machine learning models achieve better results in the field of education?

With a focus on the research question outlined in Section 1 of the Introduction, it appears that XGBoost demonstrated the highest accuracy and overall metrics, reaching 99%, making it the ideal model to use for this type of dataset.

By utilizing the six features of the dataset, which include the Portuguese and Math on second-trimester results, student satisfaction with their performance, student assessments, and total score of involvement in school, we can determine the level of risk for each student.

These findings suggest that, in practical terms, the XGBoost model is the optimal choice for predicting school failure in both Portuguese and Mathematics performance across all clusters.

5.2 Future work

Concerning the implementation of machine learning algorithms for successfully predicting students at risk of failing school, it is crucial to take action based on these insights. For future work, it would be beneficial to further investigate this study by developing a decision support system that incorporates the predictions made. By utilizing decision-making methods, this system can assist decision-makers in taking appropriate actions in this regard. Given that the problem involves multiple criteria and a predetermined set of alternatives, decision-making methods are mentioned in order to determine the optimal alternative. These methods take into account multiple criteria during the selection

process, enabling a comprehensive evaluation that considers various factors when making the final decision. The purpose of the work would involve conducting a detailed analysis of multi-criteria decision-making methods, considering that our problem relies on multiple criteria. Furthermore, it would entail implementing these methods using the original dataset along with the predictions made. Subsequently, the results obtained would serve as a guideline for decision-makers to take appropriate actions and assist students in improving their performance. The proposed architecture would be as follows:

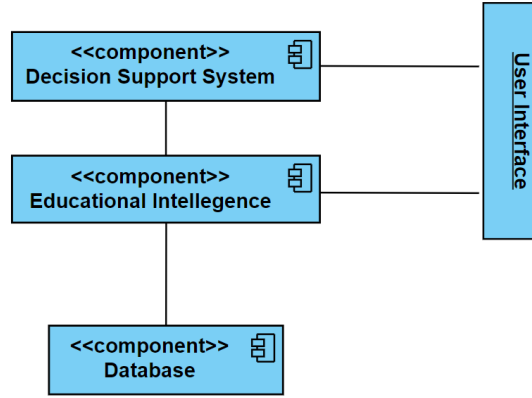


Fig. 2. Caption

The database refers to the one mentioned in section 3. The Educational Intelligence refers to the machine learning method implemented, XGBoost, as we concluded in the previous section. This method would generate a new dataset that combines the original dataset with an additional column containing the predictions obtained from the model. Finally, these results would be applied to new multi-criteria decision-making methods, which would produce a ranking list of students from higher to lower risk of failing. This list would enable decision-makers to take appropriate actions and interventions for students based on their level of risk.

6 Conclusions

The objective of this paper is to examine a study that aimed to identify the most efficient models for preventing failure in secondary education. Based on the background study, a dataset was used for the study.

After making the pre-processing data of the dataset, we had the initial 117 features, but only six were found to be significant in determining the level of risk for students.

It was applied 5 machine learning models: random forester, decision tree, XGBoost, gradient boos, and ADA Boost. After comparing the outcomes of various models to determine which approach is the most effective, we conclude that XGBoost had the best results. That is the answer to the research question.

For Futute work, and based on the level of the risk of the student, we intend to create a decision model to alert. The alert can be for teacher, parents and/or psychologist dependeing on the level of the risk.

Acknowledgements Please place your acknowledgments at the end of the paper, preceded by an unnumbered run-in heading (i.e. 3rd-level heading).

References

1. Grifo, A., & Marques, J. L. (2023). A influência da governação europeia da Educação nas políticas educativas portuguesas. *Sensos-e*, 10(1), 91-102.
2. Rashad, H. M., Aly, W. M., & Hegazy, O. F. (2013). An Educational Data Mining System for Advising Higher Education Students. *IJIE*, 7(10).
3. Rashad, H. M., Aly, W. M., & Hegazy, O. F. (2013). An Educational Data Mining System for Advising Higher Education Students. *IJIE*, 7(10).
4. Carneiro, D., Novais, P., Durães, D., Pego, J. M., & Sousa, N. (2019). Predicting completion time in high-stakes exams. *Future Generation Computer Systems*, 92, 549-559.
5. Phauk, S., Sokkhey, P., & Okazaki, T. (2020). Developing Web-based Support Systems for Predicting Poor-performing Students using Educational Data Mining Techniques. *International Journal of Advanced Computer Science and Applications*. 11. 23-32. 10.14569/IJACSA.2020.0110704.
6. Wang, J., Jiayan, T., Jiarun, L., Zhi, L., & Luo, J. (2021). Analysis of the K12 Education of United States Using Machine Learning and Data Mining Techniques. *Intelligence and Internet of Things Engineering, ICBAIE 2021*. 93-96. 10.1109/ICBAIE52039.2021.9389887.
7. Begum, S. & Padmanavar, S. (2022). Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction. *International Journal of Intelligent Engineering and Systems*. 10.22266/ijies2022.0430.29
8. Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.
9. 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.
10. Sokkhey, P., & Okazaki, T. (2020). Hybrid machine learning algorithms for predicting academic performance. *Int. J. Adv. Comput. Sci. Appl*, 11(1), 32-41.
11. Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. *International Journal of computer applications*, 36(11), 8-12.
12. Giannakas, F., Troussas, C., Krouska, A., Sgouropoulou, C., & Voyiatzis, I. (2021). Xgboost and deep neural network comparison: The case of teams' performance. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17* (pp. 343-349). Springer International Publishing.