

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-10-15

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:

Rodrigues, D., Lopes, A. L., Mauritti, R, Roque Ferreira, M. & Pintassilgo, S. (2025). Early detection of at-risk university students using machine learning: A study of model performance in evolving academic environments. In Luis Gómez Chova, Chelo González Martínez, Joanna Lees (Ed.), EDULEARN25 Proceedings (pp. 6538?6545). IATED. <https://doi.org/10.21125/edulearn.2025.1610>

Further information on publisher's website:
[10.21125/edulearn.2025.1610](https://doi.org/10.21125/edulearn.2025.1610)

Publisher's copyright statement:

This is the peer reviewed version of the following article: Rodrigues, D., Lopes, A. L., Mauritti, R, Roque Ferreira, M. & Pintassilgo, S. (2025). Early detection of at-risk university students using machine learning: A study of model performance in evolving academic environments. In Luis Gómez Chova, Chelo González Martínez, Joanna Lees (Ed.), EDULEARN25 Proceedings (pp. 6538?6545). IATED. <https://doi.org/10.21125/edulearn.2025.1610>, which has been published in final form at <https://dx.doi.org/10.21125/edulearn.2025.1610>. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

EARLY DETECTION OF AT-RISK UNIVERSITY STUDENTS USING MACHINE LEARNING: A STUDY OF MODEL PERFORMANCE IN EVOLVING ACADEMIC ENVIRONMENTS

D.M. Rodrigues¹, A. Lopes², R. Mauritti¹, M. Roque Ferreira¹, S. Pintassilgo¹

¹Iscte - Instituto Universitário de Lisboa, CIES-ISCTE (PORTUGAL)

²Iscte - Instituto Universitário de Lisboa, IT-Iscte (PORTUGAL)

Abstract

This study explores the early detection of at-risk students using machine learning (ML) models trained using historical academic data. We constructed two datasets from university enrollment records spanning the 2016/2017 to 2022/2023 academic years, focusing on first-time enrollees in their first curricular year of undergraduate and integrated master's programs. The first dataset contained information available at the start of the academic year, while the second incorporated students' first-semester performance. Each dataset was used to define two target variables: academic success and dropout risk. The tested models included Random Forest, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, AdaBoost, K-Nearest Neighbours (KNN), and Logistic Regression. To optimize ML model performance, we employed Optuna for hyperparameter tuning. The best-performing parameters for each model were then evaluated using two test datasets: the first consisting of a subset of the data of students at the start of each academic year from 2016/2017 to 2022/2023, and the second from a subset of the data for the 2023/2024 academic year, which reflects a real-world shift due to a newly implemented intervention strategy at the University aimed at providing personalized support for at-risk students. While the intervention activity itself was not included as a variable in the models, it may have influenced student success and dropout rates in 2023/2024, thus affecting the models' predictions. By comparing performance on pre- and post-intervention data, this study allows us to examine the robustness and generalizability of ML-based early warning systems in detecting at-risk students when applied to an evolving academic environment.

Keywords: Machine Learning, Early Warning Systems, Student Dropout Prediction, Academic Performance Analysis, Intervention Impact.

1 INTRODUCTION

Student success and retention are critical priorities for higher education institutions worldwide. As university enrollment grows each year, driven by increasing national demand and a surge in international students, institutions face new challenges in supporting a more diverse student population. Incoming students vary widely in socioeconomic and educational background, language proficiency, and technological readiness. Many relocate from different regions or countries, which can compound their sense of disconnection and academic stress.

Students struggling to adapt to this environment may be unlikely to seek help. Feelings of alienation or inadequacy may prevent them from accessing support services, causing academic challenges to accumulate unnoticed [1]. These unresolved difficulties can ultimately lead to course failure or withdrawal.

Early identification and intervention are therefore crucial. The sooner at-risk students are identified, the higher the likelihood of reversing negative trajectories and improving outcomes. At our university in Lisbon (Portugal), identifying such students has traditionally relied on manual, time-intensive methods conducted during the academic year. While valuable, these approaches are limited in scalability and consistency.

In this study, we propose a data-driven approach to early detection, using machine learning (ML) models trained on historical academic data. We aim to evaluate a range of algorithms and determine which provide the most accurate and generalizable predictions of two target variables: student success and dropout risk.

Furthermore, we train machine learning models to identify students who are likely to encounter academic difficulties prior to the start of the academic year. This approach relies solely on non-academic data,

excluding performance indicators such as grades, which contrast with traditional manual methods that depend heavily on early academic results to flag struggling students.

To finalize, we test the models not only on a subset of the dataset that was used for training, but also with a more recent dataset, to measure how these algorithms behave with new data that is prone to real world changes in an ever-changing environment.

By automating risk detection, our goal is to identify at-risk students as early as possible, potentially even before academic difficulties emerge, and to uncover cases that may be overlooked in manual reviews. This enables earlier, more proactive interventions from a range of available student support services, including some that students may not be aware of, ultimately contributing to improved retention and academic success.

In the following sections, we begin with a review of prior studies in the field, presented in the Literature Review section. We then describe our approach in the Methodology section, including the models trained, the target variables, and the data used. The Results section presents the performance of the trained models and discusses our key findings. Finally, we conclude with a summary of insights and future directions in the Conclusion.

2 LITERATURE REVIEW

The application of machine learning in predicting student academic performance and identifying at-risk students has gained significant attention in recent years. This reflects a growing interest in leveraging educational data to improve retention, early intervention, and personalized learning support strategies.

The adoption of ML algorithms, including Decision Trees, Random Forests, Naïve Bayes, Support Vector Machines (SVM), and ensemble methods such as AdaBoost, has significantly improved prediction accuracy [2, 3]. Furthermore, [3] underlines the importance of the preprocessing phase of the data for better models' performance and finds that the use of such algorithms "are sufficiently effective to early identify students' academic failures, and then they are useful to provide educators or teachers with relevant information to help your decisions.". For instance, [4] developed a tool to predict student failure risks in distance learning courses, resulting in a 14% reduction in dropout rates after tutoring plans were applied to the flagged students.

Several studies emphasized the importance of early warning systems. [3] evaluated various data mining techniques in introductory programming courses and found that decision trees and ensemble methods were particularly effective at early-stage prediction. Similarly, [5] used Random Forests to predict university dropouts based on administrative student data.

The effectiveness of ML models is strongly linked to the nature of the features used. A study [6] analysed different variables, and which were the most important in the outcome of different models. The most impactful variables ranged from personal data, such as a student's relationship status and social life, with socio-demographic data such as parents' employment and the student's age. However, it is also noted that academic variables, especially early performance indicators, are among the strongest predictors of success [7]. Furthermore, a study conducted in Portugal aimed at identifying the variables most strongly correlated with academic performance and dropout found that academic data consistently ranked among the most important features across all tested models [8].

Model optimization plays a vital role. The use of Bayesian optimization for hyperparameter tuning, as demonstrated by [9], can significantly enhance model performance, especially when dealing with imbalanced datasets.

While most models perform better when academic performance data is available, some studies also explore predictions based solely on enrollment and demographic data. These approaches are crucial for interventions at the start of the academic year, before grades are available [2].

Overall, the literature consistently supports the use of ML for predicting at-risk students, especially when integrated into scalable early-warning systems. However, it is also important to consider that model performances can be affected by evolving contexts, such as the introduction of new student support initiatives, which in turn can influence the academic outcome of students, potentially reducing the performance of the trained models.

3 METHODOLOGY

In this study we used institutional academic records from a public university in Lisbon (Portugal), covering academic years from 2016/2017 to 2023/2024. For each year two datasets were constructed, the first used only student enrollment data while the second also contained academic data related to the students' first semester performance. All datasets focus exclusively on first-time enrollees in their first curricular year of undergraduate and integrated master's programs.

The study aims to predict two binary classification tasks which are academic success and dropout risk. Students are considered academically successful if they are eligible to progress to the 2nd curricular year by the end of the respective academic year. Dropout is defined as the failure to enroll in the following academic year.

It is also important to point out that in the year 2023/2024 a new implemented intervention strategy at the University was created [10], aimed at providing personalized support for at-risk students, as well as new welcoming strategies to result in a better integration of new students. These are some of the many changing factors in the real world that may influence the results of students and in turn can also affect the models' predictions.

Since the new personalized interventions added in the 2023/2024 academic year are targeted at-risk students, these interventions may have improved the performance of students who would otherwise have struggled, and therefore, some at-risk individuals in the 2023/2024 dataset may not exhibit academic failure. This is important to keep in mind, since it can lead to the model not identifying students with a similar risk profile, although they would benefit from personalized intervention. For this reason, we decided to not use the data from 2023/2024 to train the models, instead, all models will be trained with data from 2016/2017 until 2022/2023 (inclusive) and they will be tested with 2 different datasets, one is a subset of data from the same period (until 2022/2023) and another test will use the data from 2023/2024.

3.1 Datasets Variables

We used two different kinds of data, enrollment data, available from the moment students enroll in the university, and academic data, available after the first semester is over.

For the categorical variables in each dataset, any missing values were replaced with None while missing numerical data was replaced with -1. The categorical variables were also one hot encoded, a technique that transforms each category into a separate binary column (e.g. The column 'nationality' is split into new columns, one for each nationality in the dataset, with the values True or False for each of them), allowing models to interpret non-numeric data. While for numerical variables we applied a standard scaler, which standardizes features and scales them to unit variance, ensuring all numeric features are on the same scale.

Enrollment data:

- Categorical variables:
 - Gender
 - Nationality
 - Second nationality
 - Home country
 - Displaced (Boolean - student whose home is far from university)
 - students are displaced if one of the following conditions is met: they come from outside of Portugal mainland (international student or islands); their place of residence is over 50km from the University; Expected arrival time to University exceeds 60 minutes.
 - Admission category (e.g. general admission, special needs, international, etc.)
 - Admission phase
 - Enrollment date
 - Enrollment mode (Full-time or Part-time)
 - Enrollment status (Enrolled, asked to leave, dropout, concluded, etc.)
 - In debt (Boolean - student in debt with the university)

- Scholarship result (granted, denied, not asked),
- Degree
- School (The university is divided into several schools)
- Campus (The university has two campuses)
- Statutes (e.g. special needs, working student, etc.)
- Numerical variables:
 - Age
 - Admission grade
 - Enrollment option number (students can apply to 6 universities, this indicates where the university was in the list, 1 - first choice, 6 - last choice)

Academic data:

- Numerical variables:
 - Enrolled ECTS first semester
 - Concluded ECTS first semester
 - Ratio concluded/enrolled ECTS first semester
 - Not assessed (NA) ECTS first semester
 - Ratio not assessed (NA)/enrolled ECTS first semester

4 MACHINE LEARNING MODELS

We implemented the machine learning models using the Python library scikit-learn (sklearn). We first created a sklearn pipeline with all the necessary preprocessing steps, such as handling missing values, features scaling and categorical encoding. This allowed us to ensure that the datasets always undergo the same preprocessing in order to not influence the results.

The models trained and tested with the sklearn library present in this paper are Random Forest, Support Vector Machine (SVM), Naïve Bayes, Decision Tree, AdaBoost, K-Nearest Neighbours (KNN), and Logistic Regression. These are widely used and strong performing models as reported in related works. Other promising options that have shown a big potential such as neural networks (NN) were not explored in this study due to the amount of data needed in order to effectively train a NN.

For all possible combinations of model, dataset and target variable, hyperparameter tuning was performed using the Optuna framework, which applies an efficient search strategy based on Bayesian optimization. A total of 150 trials were done for each combination, which helped ensure that each model was then evaluated using its best-performing configuration.

In order to evaluate the model's hyper parameterization, we looked at 4 different metrics, also calculated via the sklearn library: accuracy_score, recall_score, f1_score and roc_auc_score.

- Accuracy - Measures the proportion of total correct predictions made by the model across all classes.
- Recall - Measures the proportion of positive cases that were correctly identified.
- F1 Score - Harmonic mean of Precision and Recall, balancing the results when the classes are imbalanced.
- ROC AUC (Receiver Operating Characteristic – Area Under the Curve) - Measures how well the model distinguishes between the two classes).

4.1 Training and Testing ML Models

As discussed before, to train all the ML models we used only data from 2016/2017 until 2022/2023, separated into two different datasets, one with just the enrollment data (start of the year), and another with the academic data added (end of first semester). Each model was then trained using the best performing configurations from the Optuna trials, with all the data from 2016/2017 to 2022/2023, passing through the same pipeline as the one the data went through before hyper parameterization on Optuna.

For testing, we then used those trained models and tested each with a subset of that same data until 2022/2023, and another with the data from 2023/2024, meaning that each model was tested with two

different datasets. The same four metrics as before were calculated for each test, to analyse the performance of each model.

5 RESULTS

In this section we present the results of all the tests performed to each model for both target variables (success and dropout) and for both the tested datasets (start of the year and end of the 1st semester), discussing each one individually and then comparing the results, in order to analyse how real-world changes can affect the performance of these ML algorithms.

5.1 2016/2017 to 2022/2023 Testing Dataset

Table 1. Models' metrics results for 2016-2023 testing dataset.

Target	Dataset	ML Model	Accuracy	Recall	F1-Score	Roc Auc
Success	Start of the year	Random Forest	94	99	96	89
		SVM	79	98	87	62
		Naïve Bayes	74	86	83	63
		Decision Tree	99	99	100	99
		AdaBoost	78	95	86	63
		KNN	82	95	88	71
		Logistic Regression	78	96	87	63
	End 1 st semester	Random Forest	94	98	96	91
		SVM	95	99	96	91
		Naïve Bayes	89	92	82	86
		Decision Tree	100	100	100	100
		AdaBoost	94	97	96	91
		KNN	94	99	96	89
		Logistic Regression	94	97	96	91
Dropout	Start of the year	Random Forest	97	100	98	90
		SVM	85	100	92	56
		Naïve Bayes	80	90	88	59
		Decision Tree	100	100	100	100
		AdaBoost	85	99	92	57
		KNN	87	98	92	64
		Logistic Regression	85	99	92	57
	End 1 st semester	Random Forest	95	99	97	88
		SVM	94	99	97	85
		Naïve Bayes	89	92	93	84
		Decision Tree	100	100	100	100
		AdaBoost	94	98	96	85
		KNN	94	99	96	85
		Logistic Regression	94	99	97	85

On Table 1 we can see the results of testing each model for the date period of 2016/2017 to 2022/23. Overall, the Decision Tree was the model that performed best in all metrics, closely followed by Random Forest which also had a strong performance. The Naïve Bayes models had the worst performance in all tests. On the remaining models there is a significant improvement on the results of the end of the 1st semester, indicating the reliability of academic data to help to correctly predict the at-risk students.

Regarding the start of the year models' performances, the ROC AUC values of all the models except Random Forest and Decision Tree, are concerning, since they all scored between 56 and 71, indicating less reliable class separation despite the high recall scores.

With the addition of the academic data at the end of the 1st semester datasets, SVM, AdaBoost, KNN and Logistic regression all had similar and good performing scores.

5.2 2023/2024 Testing Dataset

Table 2. Models' metrics results for 2023/2024 testing dataset.

Target	Dataset	ML Model	Accuracy	Recall	F1-Score	Roc Auc
Success	Start of the year	Random Forest	79	97	88	58
		SVM	78	99	87	54
		Naïve Bayes	77	87	85	66
		Decision Tree	68	77	78	57
		AdaBoost	80	97	88	60
		KNN	77	91	86	59
		Logistic Regression	72	78	71	65
	End 1 st semester	Random Forest	93	97	95	88
		SVM	93	98	95	86
		Naïve Bayes	87	86	91	88
		Decision Tree	71	69	79	75
		AdaBoost	89	99	93	76
		KNN	93	98	95	87
		Logistic Regression	22	00	00	-
Dropout	Start of the year	Random Forest	85	96	92	56
		SVM	86	100	92	-
		Naïve Bayes	80	88	88	59
		Decision Tree	71	77	82	55
		AdaBoost	86	98	92	53
		KNN	84	95	91	54
		Logistic Regression	86	100	92	-
	End 1 st semester	Random Forest	89	92	93	80
		SVM	86	100	92	-
		Naïve Bayes	78	78	86	79
		Decision Tree	86	90	92	76
		AdaBoost	91	98	85	72
		KNN	89	98	94	67

		Logistic Regression	91	98	95	71
--	--	---------------------	----	----	----	----

In Table 2, we present the results of testing all models on the 2023/2024 dataset. In contrast to the previous testing period (2016–2023), no single model clearly outperformed the rest. Notably, models such as Decision Tree and Random Forest, which had previously achieved top performance, did not maintain their dominance in this new test.

For the start of the year models, the ROC AUC metric was consistently low for all models across both target variables, ranging from 53 to 66, except for SVM and Logistic Regression models for the Dropout target, where no value could be calculated since these models always predicted the same class. The Decision Tree model who previously had the top performance, had the lowest overall performing model with the 2023/2024 data. The best performing models were Naive Bayes and AdaBoost with the best overall metrics.

When it comes to the end of the 1st semester models, the performance of most models improved thanks to the addition of the academic records. For the Success target, Random Forest, SVM and KNN all had similar and strong metrics values, while for the Dropout target variable, Random Forest, AdaBoost and Logistic Regression were the ones with the best performances.

This second test, which used more recent data than the testing data, show a clear decline in the models' performances, mainly in the best performing models in the first test. Nevertheless, most models still achieved competitive results and demonstrated their utility in identifying at-risk students, even in the presence of significant environmental changes.

In addition to model accuracy and general performance, it is also important to consider the number of at-risk students flagged by each model. When two models achieve similar performance metrics, the one that identifies a significantly higher number of at-risk students may require more institutional resources to intervene effectively (e.g. for the start of the year models predicting success, the number of flagged students ranged from 46 to 501). This can strain support systems and risk diminishing the quality or timeliness of each intervention. Therefore, selecting a model must balance predictive performance with practical feasibility and intervention capacity, ensuring that the system remains both accurate and sustainable in an operational context. To address this, a manual review of the flagged cases, can help refine the intervention strategy. This process supports the identification of priority cases and may also reveal false positives, allowing advisors to focus resources where they are most needed while building trust in the system's outputs.

6 CONCLUSIONS

As expected, the machine learning models performed strongly when academic data from the first semester was included, highlighting the predictive power of early academic performance. Nevertheless, several models also showed promising results despite facing a slight decline in performance, when using only enrollment data available at the start of the academic year. This demonstrates the potential of early, data-driven risk detection systems that do not rely on academic outcomes, enabling timely interventions even before students begin their course.

The introduction of institutional interventions in the 2023/2024 academic year provided a unique opportunity to evaluate the robustness of these models under real world change. While some previously top performing models experienced a drop in performance under these new circumstances, many maintained acceptable metrics values, supporting their generalizability and relevance across evolving academic environments, but also raising the importance of testing these models and if the need arises, to fine tune them with new data in later years.

For future work, we aim to explore the introduction of additional data sources that were not available during this investigation. This includes socio-demographic variables such as parents' educational background and occupations, as well as academic engagement data from platforms like Moodle, which can provide insights into students' access to and interaction with course materials, including lecture slides, exercises, and past exams. These additional data points may have a meaningful correlation with the target variables of academic success and dropout risk, potentially enhancing the predictive power of the models.

ACKNOWLEDGEMENTS

This work was financially supported by Project In_Iscte2 Espaço para Crescer II, Call 05/C06-i07/2023, funded by the Portuguese Recovery and Resilience Program (PPR), The Portuguese Republic and The European Union (EU) under the framework of Next Generation EU Program.

REFERENCES

- [1] White M, Canning EA. "Examining active help-seeking behavior in first-generation college students". *Social Psychology of Education*. Vol. 26, pp. 1369-1390, 2023, doi: 10.1007/s11218-023-09794-y.
- [2] R.Z. Pek, S.T. Özyer, T. Elhage, T. ÖZYER, R. Alhajj, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure", *IEEE Access*, vol. 11, pp. 1224-1243, 2023, doi: 10.1109/ACCESS.2022.3232984.
- [3] E.B. Costa, B. Fonseca, M.A. Santana, F.F. Araújo, J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", *Computers in Human Behavior*, vol. 73, pp. 247-256, 2017, doi: 10.1016/j.chb.2017.01.047.
- [4] C. Burgos, M.L. Campanario, D. Peña, J.A. Lara, D. Lizcano, M.A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout", *Computers & Electrical Engineering*, vol. 66, pp. 541-556, 2018, doi: 10.1016/j.compeleceng.2017.03.005.
- [5] Behr, Andreas & Giese, Marco & Tegum, Herve & Theune, Katja, "Early Prediction of University Dropouts – A Random Forest Approach", *Jahrbücher für Nationalökonomie und Statistik*, vol. 240, 2020, doi: 10.1515/jbnst-2019-0006.
- [6] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, Y. El Alloui, "A Multiple Linear Regression-Based Approach to Predict Student Performance", *AI2SD 2019*, vol 1102, pp: 9-23, 2020, doi: 10.1007/978-3-030-36653-7_2
- [7] A. Hellas, P. Ithantola, A. Petersen, V.V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S.N. Liao, "Predicting academic performance: a systematic literature review.", *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE 2018 Companion)*. Association for Computing Machinery, New York, NY, USA, pp.175–199, 2018, doi: 10.1145/3293881.3295783
- [8] V. Realinho, J. Machado, L. Baptista, M. V. Martins, "Predicting Student Dropout and Academic Success", *Data*, vol. 7(11), 2022, doi:10.3390/data7110146
- [9] S. Albahli, "Efficient hyperparameter tuning for predicting student performance with Bayesian optimization", *Multimedia Tools and Applications*, vol. 83, pp: 52711–52735, 2024, doi: 10.1007/s11042-023-17525-w
- [10] A.L. Lopes, R. Mauritti, S.C. Martins, S. Pintassilgo, C. Matias, S. José (2024) "DIGITAL TOOLS FOR THE PREVENTION OF DROPOUT AND ACADEMIC FAILURE: A CASE STUDY OF A PORTUGUESE UNIVERSITY", *EDULEARN24 Proceedings*, pp. 7092-7098, doi: 10.21125/edulearn.2024.1678