Advances and Applications in Mathematical Sciences
Volume 23, Issue 4, February 2024, Pages 279-299

CUSTOMIZED LEARNING TO PREDICT STUDENT DROPOUT

ELENA BALLANTE[1*], FARAH NAZ[2], SILVIA FIGINI[1] and SIMONE GERZELI[1]

[1*]Department of Political and Social Sciences
University of Pavia, via Corso Strada Nuova 65
Pavia, 27100, Italy

[2]Department of Mathematics
University of Pavia, via Ferrata 5
Pavia, 27100, Italy

Abstract

Student dropouts are a fundamental problem for higher education institutions and constitute an area of extensive research. This paper presents the application of machine learning techniques for the early identification of students who are at risk of dropping out, enabling institutions to design effective prevention strategies and implement retention activities. The paper analyzes real data provided by the University of Pavia, including information about students enrolled in the first year of all the courses. The use of data that are automatically registered at the enrollment is the main strength of this work, making the method feasible without an ad hoc data collection. In fact, the computational approach proposed in this paper can be easily generalized for the application to other academic institutions to predict the likelihood of student attrition and implement targeted interventions for student dropouts. The main novelty of this paper is the development of a monitoring dashboard that includes the results obtained by machine learning models and is available for counseling offices as well as student offices to underline the real-time behavior of students at risk, thus planning retention activities.

1. Introduction

Student dropout is among the most challenging problems affecting academic institutions. It is a complex phenomenon with critical consequences both for a student's career and for the academic organization itself (C. Barra,

2020 Mathematics Subject Classification: 62P25.
Keywords: University careers, Dropouts, Machine Learning, Retention, Educational data.
Corresponding author; E-mail: elena.ballante@unipv.it
Received October 23, 2023; Accepted November 22, 2023

R. Lagravinese and R. Zotti [1]). A dropout is a potentially devastating event in the life of a student, and it also negatively impacts the university from an economic and reputational point of view (M. Jadrić, Ž. Garača and M. Čukušić [2]). The dropout rate for higher education institutions has encouraged researchers to present a wide range of methods to predict at-risk students, with the main objective of providing timely information that enables tutors, professors, and psychologists to select the most effective treatments to reduce student dropout (F.B. Rinaldi et al. [3]). Furthermore, the containment of student dropout is a metric used by legislators, accreditation agencies, and governing bodies to evaluate and give resources to universities. Providing students with remedial assistance at the right time has often proven an effective method to reduce student dropout, thus, the identification of students that require this type of support is mandatory.

Machine learning models coupled with domain knowledge provided by a variety of stakeholders, such as educators, counselors, advisors, and other staff members, can identify students who are "at risk". The identification of predictors that can help in this predictive task is crucial to obtaining an accurate model that can help control the phenomenon.

The investigation of student dropouts is an important area of research because of the significant impact it has on students, universities, and society as a whole. According to (S. Herzog [4]) high dropout rates in higher education prevent students from achieving their educational goals and represent a waste of resources for universities and society. Additionally, dropout rates can be an indication of broader issues within the education system, including insufficient academic preparation, inadequate social support, or financial barriers (H. Williams and N. Roberts [5]). By investigating the factors that contribute to student dropouts, universities can gain a better understanding of the challenges faced by their students and develop interventions to improve student retention and success (J.G., Piepenburg and L. Fervers [6]). Finally, such efforts can help to create a more equitable and effective higher education system that benefits students, universities, and society as a whole (N. Rotem, G. Yair and E. Shustak [7]). Statistical techniques have been predominantly employed to predict student dropout in a wide range of educational contexts, academic environments, and theoretical frameworks of the analysis (T. Zajac et al. [8]).

The primary aim of this paper is to use machine learning models to predict student dropout and to illustrate, for each course at the University of Pavia, the group of students who are at risk. The results obtained using machine learning models are depicted using a monitoring dashboard developed in-house and available to plan retention activities by a team of experts that includes psychologists, possessors, and tutors. The dashboard developed in-house is a robust tool for the early identification of students' difficulties in a specific course in which they are enrolled. In our opinion, data analytics coupled with non-cognitive factors such as motivation, self-efficacy, perseverance, career goals, self-regulated learning strategies, and intention to continue their studies play a vital role in determining academic success or, conversely, the likelihood of dropping out (C. Bargmann, L. Thiele and S. Kauffeld [9]). Universities should focus on fostering students' intrinsic motivation and providing career counseling to increase career decision marking and reduce the likelihood of dropouts (K. Cidlinska et al. [10]). Our paper also investigates which kind of data could improve the effectiveness of the results by predicting the probability of dropout with the data automatically recorded at enrollment, with the possibility of also including information about the participation of students in orientation events organized by the university. The data at hand provided by the University of Pavia includes 4574 students enrolled in the first year of all the courses offered by the University of Pavia.

The methodological approach described in the paper, coupled with the monitoring dashboard, is general enough to be delivered to other university institutions around the world. Different metrics are employed to measure the prediction models' performance and to assess the accuracy and validity of the proposed algorithms, including cross-validation exercises to apply to real data. The research paper is structured as follows: Section 2 reports concerning student dropouts in universities. The dataset and the methodology are presented in Section 3 including the description of a dynamic dashboard that can help academic institutions make informed decisions. Section 4 discusses the results at hand and proposes future ideas for research based on more extensive databases.

2. Literature Review

High dropout rates in universities are a serious problem among undergraduates. Several papers recently addressed the prediction of student dropout using different machine learning and statistical techniques. The problem of dropouts in Italian universities was approached from several points of view, including an aggregated level of cultural information (E. Ripamonti and S. Barberis [11]), in a longitudinal framework (S. Meggiolaro, A. Giraldo and R. Clerici [12], [13]), with neural network technique (F. Agrusti [14]).

In (C. Marquez-Vera, C.R. Morales and S.V. Soto [15]), predicting school failure and dropout in a high school in Spain is described using data mining techniques. The study found that both algorithms were effective in predicting school failure and dropout. The authors also identified several key factors that were associated with a higher risk of school failure and dropout, including low academic achievement, truancy, and being male. In the case of university-level education, (D. Delen [16]) proposed a methodology based on machine learning models to analyze five years of institutional data. They investigated the reasons behind freshman student attrition, revealing that educational and financial variables are among the most important predictors of the phenomenon. The authors explained that a balanced dataset produced better prediction results than an unbalanced dataset. This point is widely discussed in (D. Thammasiri [17]) where different data balancing techniques are compared to improve predictive accuracy in the minority class while maintaining satisfactory overall classification performance.

D. Rodríguez-Gómez et al. [18] presented a study on using predictive modeling to identify students at risk of poor academic outcomes in a large public university in the United States. The authors developed a logistic regression model based on student demographic, academic, and financial aid data to predict the likelihood of poor academic outcomes, defined as low GPA, academic probation, or academic dismissal. The study found that the predictive model had good accuracy in identifying at-risk students. The authors also identified several key risk factors for poor academic outcomes, including a low high school GPA, enrollment in remedial courses, and receiving low amounts of financial aid.

A. Sarra, L. Fontanella and S. Di Zio [19] analyzed data related to 561 students enrolled in courses at an Italian university. The data were collected as online questionnaires (through the computer-assisted web interviewing method) and investigated different aspects of the student's academic lives, focusing on psychological factors. Also, the response variable was related to the intention to dropout instead of objective data about observed dropouts. Data were analyzed by applying Bayesian profile regression, a technique that incorporates concepts from latent class analysis into a regression framework.

F. Del Bonifro et al. [20] developed recently a tool that allows estimating the risk of quitting an academic course based on real data of students from eleven schools of a major Italian university. The decision support system considers the dataset's statistical composition (highly unbalanced for the classes) and provides predictions at the moment of student enrollment. They aim to integrate this tool into a more general monitoring system useful for university governance. Another research work on student retention in universities was proposed in.

(M. Kadar et al. [21]) The novelty of this research is the nature of the input data used for predicting the phenomenon. The approach exploits data acquisition by webcams, eye-trackers, and other similar devices in the context of an IoT class. Based on these data, it is possible to perform emotion analysis and detection for the students in the room, which will then be exploited to predict dropout. Despite the novelty of this approach, the data acquisition process is not so easy to extend to other universities.

Also, P. Perchinunno, M. Bilancia and D. Vitale [22] approached the problem of dropouts in higher education. They analyze aggregate data from an agency for the evaluation of the university and research system to estimate the dropout problem at a national level. Then analyze the data collected from the University of Bari Aldo Moro at the individual level. They analyze dropout events in the first and second years of university, including information about the first-year careers of the students. They applied parametric and non-parametric models to evaluate the predictive power of the collected data and investigate the most important features. With respect to the literature review, our contribution is two-fold: firstly, a monitoring dashboard is developed to share the results for a real-world application; secondly, the improvement of the data including the orientation events and

the variables related to the student history before or at the enrollment (score at Lyceum) to develop a predictive machine learning approach that does not require ad hoc data collection.

3. Empirical Analysis

The analysis focuses on two different sources of data from the University of Pavia (one of the major universities in the North of Italy): aggregated data about students enrolled in the first year from 2015 to 2019 aimed at the development of a monitoring dashboard, and detailed data about each student enrolled in the first academic year 2019/2020 used for the training phase of a predictive algorithms to forecast student dropout.

3.1 Monitoring dashboard

To investigate some recurrent trends and behaviors in the dropout student dataset, a dynamic dashboard has been developed with the help of colleagues. The variables included in the dashboard are the university department, course, and the following variables: age of enrollment, gender, type of high school, area, tax exemption, and participation in orientation events.

Figure 1 shows the trends of first-year dropout rates for each area (Engineering, Humanistic, Law-Economic-Politic, Medical, Scientific) at the University of Pavia. In the same figure, the absolute number of enrollments has been added. Note that the dropout rates reported for the medical area are much lower than the other areas. This is mainly due to the fact that the admission process is very different with respect to the other areas (there is an admission test planned at the national level). This justifies the exclusion of students belonging to the medical and biological areas from the predictive analysis. Comparing dropout rates to the number of enrollments per year, we observe that where the number of new enrollments is high, the dropout rates seem to show greater value. The reason for this observed correlation between enrollment rates and dropout rates may be attributed to several factors, such as different student features (in terms of demographics), increased student support services, and a strong institutional reputation.

Dropout rates are described more in detail in Table 1, where, for each department, the average dropout rates, as well as enrollments for the last 5

years, are listed coupled with the lower and upper limits for the individual output for both observations. The most critical department is biology. The reason behind this is the enrollment of those students whose final aim is to study medicine, but they are blocked at the admission test and decide to attend biology for one year, temporarily, but they move to medicine the following year. Of course, this type of dropout cannot be avoided since students do not leave biology because of difficulties in their study plans.

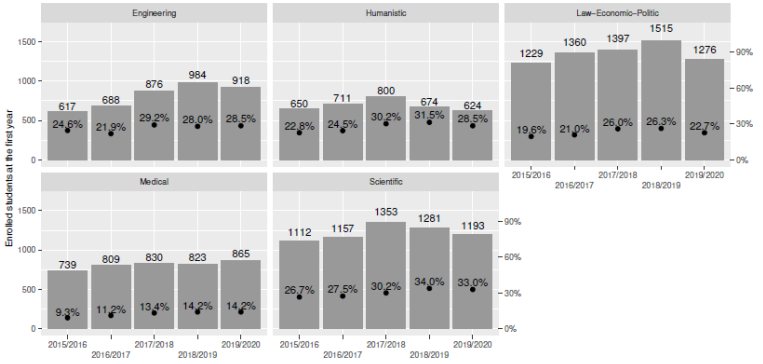


Figure 1. First-year dropout rates for each study area of the University of Pavia coupled with absolute numbers of first-year enrollments (Barplot included in the monitoring dashboard).

Table 1. Average first-year dropout rates and enrollments for the last 5 years in the University of Pavia.

Department/course	Dropout rates	Enrollments
CIVIL ENGINEERING AND ARCHITECTURE	26.2% (± 6.0)	154 (± 26)
INDUSTRIAL AND INFORMATION ENGINEERING	26.6% (± 2.7)	662 (± 135)
HUMANISTIC STUDIES	30.5% (± 5.2)	484 (± 74)
MUSICOLOGY AND CULTURAL HERITAGE	29.7% (±5.3)	86 (± 9)
LAW	27.9% (± 6.5)	256 (± 18)

ECONOMICS AND MANAGEMENT SCIENCES	17.8% (± 2.3)	539 (± 20)
POLITICAL AND SOCIAL SCIENCES	26.0 % (± 3.5)	560 (± 109)
MEDICINE	1.4 % (± 1.4)	151 (± 9)
MOLECULAR MEDICINE	6.2 % (± 1.9)	106 (± 9)
HEALTH SCIENCES	17.0% (± 4.1)	410 (± 32)
BRAIN AND BEHAVIORAL SCIENCES	13.1% (± 2.5)	165 (± 9)
DIAGNOSTIC PAEDIATRIC CLINICAL AND SURGICAL SCIENCE	17.0% (± 3.6)	103 (± 14)
BIOLOGY	37.0 % (± 3.7)	444 (± 21)
CHEMISTRY	18.9 % (± 4.1)	91 (± 4)
PHYSICS	20.7 % (± 1.9)	74 (± 13)
MATHEMATICS	27.3% (± 10.2)	58 (± 15)
PHARMACEUTICAL SCIENCES	26.0% (± 6.1)	373 (± 15)
EARTH AND ENVIRONMENTAL SCIENCES	32.6% (± 5.4)	179 (± 53)

A predictor of the dropout rate of students at the university is the type of diploma obtained during high school. Some high schools in Italy are well organized to prepare students for university life, such as the Lyceum. As we can see in Table 2, retention rates differ a lot in the categories listed. The most problematic students are those from professional high schools, where students are usually more prepared for working life than for university. The results of the dashboard are used in section 3.2.

Table 2. Average first-year dropout rates and enrollments of the last 5 years in the University of Pavia.

Type of high school diploma	Dropout rates	Enrollments
FOREIGN	19.6% (± 5.9)	214 (± 60)
TEACHER TRAINING	26.6% (± 4.1)	378 (± 41)
LYCEUM	21.0% (± 3.2)	2706 (± 124)
TECHNICAL	29.2% (± 1.5)	1076 (± 140)
PROFESSIONAL	34.4 % (± 4.5)	315 (± 40)
NOT PROVIDED	27.8 % (± 5.3)	208 (± 75)

3.2 Predictive algorithms

The dataset used for the training of the predictive models considers only the first-year students enrolled at the University of Pavia (freshmen) in the academic year 2019/2020, excluding all the courses related to the medical and biological areas, for a total of 3223 students. The information available in the dataset is demographics (age, sex, residential code, nationality), educational schools (high school score, high school graduation year, type of high school), enrollment-related information (type of enrollment in the course, age of enrollment, examined knowledge gap), information about tax payment, department, and course. Information about participation in orientation events before enrollment has also been added to the classical input variables used for the prediction of freshmen attrition. This type of information is derived from the student archive, which records all the subscriptions and registrations to university orientation events, both for students at the university and for high school students.

Different machine learning models have been compared to understand if the input variables collected are sufficient and useful to predict a first-year dropout. In the first step, the predictive models have been trained on the first block of variables, excluding the information about participation in orientation events. In the second step, participation in orientation events has been included in the predictor set to verify that this type of information can improve predictive performance. All the models have been compared using 10-fold cross-validation, and the AUC index was computed for each model to take into consideration the unbalance of the data with respect to the target variable (dropout or not dropout).

The a priori probability of dropout in our dataset is 29%.

It is evident from the substantial amount of research published on the topic and its significant socio-economic impact that the prediction of university dropout has garnered considerable attention within the scientific community. In our study, we have compared different techniques proposed in machine learning to predict dropout. The models under comparison are Naive Bayes (NB), Logistic Regression (LOGR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost (ADABOOST), and Linear Mixed Model (MIXED). However, the identification of a technique that outperforms others in terms of prediction accuracy is contingent on various factors, including the context, the characteristics of the data, and the technique itself.

In the context of predicting student dropout, a decision tree (DT) is a method that utilizes a hierarchical structure of conditions. According to (D. Heredia, Y. Amaya and E. Barrientos [23]), this technique is used due to its flexibility in handling numerical and categorical data, monotonic transformations of explanatory variables, and ease of interpreting results. DT also offers better accuracy rates compared to other methods. Support vector machines (SVMs) are frequently used in the literature to predict student dropout in online courses due to their effectiveness in solving classification problems (J. Liang et al. [24]). SVMs are often preferred due to their simplicity and ease of understanding. According to [4], logistic regression analysis can better identify significant predictors of each outcome. Naive Bayes works by calculating the probability of each class for a given input, based on the presence or absence of certain features in that input, even with large datasets (L. Paura and I. Arhipova [25]). Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their outputs to produce a final prediction. The algorithm is known for handling large datasets and variables with high correlation. According to (S. Sivakumar, S. Venkataraman and R. Selvaraj [26]), after selecting important features, Random Forest is an effective algorithm for predicting student dropout and can be used by institutions to take appropriate measures to prevent dropout. The machine learning algorithm AdaBoost works by combining features with high predictive power to get more accurate predictions (Y. Chen, A. Johri and H. Rangwala [27]). Whereas, Linear Mixed

Model is a statistical model that is used to analyze data that has both fixed and random effects. In the context of predicting student dropout, a Linear Mixed Model can be used to model the relationships between student characteristics, course characteristics, and the likelihood of dropout over time.

The results of the analysis are presented in terms of confusion matrices and accuracy in Table 3 and Table 4, where the predictive models were trained without incorporating information on event participation. The AUCs obtained by resorting to 10-fold classification models are reported in Table 5.

Table 3. Confusion matrix using 10-fold cross validation

		NB		LOGR		SVM		DT	
		No	Yes	No	Yes	No	Yes	No	Yes
Confusion Matrix	No	2254	79in 2232	11	2182	151	2257	76	
	Yes	814	76	758	13have 746	144	805	85	
Per-class accuracy		73.47%	49.03%	74.65%	56.65%	74.52%	48.81%	73.71%	52.79%
Overall Accuracy		72.29%		75.45%		72.17%		72.67%	

The results are quite good, but there is no outstanding model both in terms of accuracy and AUCs.

The higher performances are reached by Logistic Regression and AdaBoost, so we will present results from both of these models, exploiting the advantages of parametric and non-parametric models.

The paper also investigates if the orientation events reduce student dropout by training the predictive models first without incorporating information on event participation and then including this variable in the predictor set to verify that this type of data can improve predictive performance. The results of machine learning models after incorporating the event variable are reported in Table 6, Table 7, and Table 8.

Table 4. Confusion matrix using 10-fold cross validation.

		RF		ADABOOST		MIXED	
		No	Yes	No	Yes	No	Yes
Confusion Matrix	No	2024	309	2176	157	2241	92
	Yes	617	273	702	188	780	110
Per-class accuracy		76.64%	46.91%	75.61%	54.49%	74.18%	54.46%
Overall Accuracy		71.27%		73.35%		72.94%	

Table 5. Mean AUC and minimum and maximum value for each model computed through 10-fold cross-validation

Model	AUC	(min, max)
Naive Bayes	0.630	(0.564, 0.684)
Logistic Regression	0.672	(0.613, 0.733)
SVM	0.640	(0.589, 0.684)
Decision Tree	0.563	(0.500, 0.623)
Random Forests	0.662	(0.617, 0.716)
AdaBoost	0.686	(0.625, 0.752)
Mixed Logistic model	0.671	(0.612, 0.741)

Table 6. Confusion matrix using 10-fold cross-validation (Model including orientation events).

		NB		LOGR		SVM		DT	
		No	Yes	No	Yes	No	Yes	No	Yes
Confusion Matrix	No	2244	89	2228	105	2177	156	2257	76
	Yes	809	81	759	131	738	152	805	85
Per-class accuracy		73.50%	47.65%	74.59%	55.55%	74.68%	49.35%	73.71%	52.80%
Overall Accuracy		72.14%		73.19%		72.26%		72.67%	

Table 7. Confusion matrix using 10-fold cross-validation (Model including orientation events).

		RF		ADABOOST		MIXED	
		No	Yes	No	Yes	No	Yes
Confusion Matrix	No	2037	296	2176	157	2240	93
	Yes	629	261	709	181	775	115
Per-class accuracy		76.41%	46.86%	75.42%	53.55%	74.30%	55.29%
Overall Accuracy		71.30%		73.13%		73.07%	

On the basis of the data at hand, we observe that the results are essentially stable with respect to the previous analysis.

Table 8. Mean AUC and 95% confidence interval for each model computed through 10-fold cross-validation (Model including orientation events).

Model	AUC	(min, max)
Naive Bayes	0.633	(0.566, 0.681)
Logistic Regression	0.674	(0.614, 0.728)
SVM	0.652	(0.611, 0.695)
Decision Tree	0.563	(0.500, 0.623)
Random Forests	0.662	(0.630, 0.720)
AdaBoost	0.681	(0.627, 0.752)
Mixed Logistic model	0.672	(0.614, 0.736)

The output of the Logistic Model trained on the entire dataset is reported in Table 9, where the p-values of significant variables are written in bold.

Table 9. Logistic regression model trained on the entire dataset

	Estimate Std.	Error	*z* value	*p* value
(Intercept)	-1.96	0.98	-2.01	0.04
Age enrollment	0.10	0.04	2.61	**0.01**
Gender M	0.08	0.09	0.86	0.39
Type of High School Liceo	0.13	0.28	0.46	0.65
Type of High School Professional Inst	0.72	0.31	2.35	**0.02**
Type of High School Technical Inst	0.74	0.28	2.68	**0.01**
Type of High School Unknown	0.85	0.33	2.59	**0.01**
High School Score	-0.02	0.00	-5.53	**0.00**
Dep Physics	0.53	0.44	1.22	0.22
Dep Law	1.58	0.41	3.89	**0.00**
Dep civil engineering-architecture	0.88	0.37	2.37	0.02
Dep industrial and information engineering	0.33	0.31	1.09	0.27
Dep mathematics	0.70	0.43	1.63	0.10
Dep musicology and cultural heritage	0.54	0.40	1.33	0.18
Dep drug science	2.31	0.44	5.27	**0.00**
Dep earth and environment science	0.54	0.38	1.41	0.16
Dep economics science	-0.34	0.31	-1.11	0.27
Dep p-valueal and social sciences	-0.03	0.31	-0.08	0.93

Dep humanistic studies	0.85	0.32	2.66	**0.01**
Type of course single cycle master	-1.35	0.31	-4.31	**0.00**
Type of enrollment local programming	0.12	0.14	0.80	0.42
Type of enrollment national programming	0.00	0.61	0.00	1.00
Area Abroad	0.28	0.79	0.35	0.73
Area Milan	0.25	0.34	0.75	0.45
Area North Italy	0.07	0.33	0.22	0.83
Area Pavia	0.04	0.34	0.12	0.90
Area South Italy	0.54	0.35	1.53	0.12
Tax exemption Yes	-0.49	0.09	-5.61	**0.00**
Events Yes	-0.27	0.11	-2.39	**0.02**
Age difference	-0.07	0.04	-1.61	0.11

The variables explain and predict dropouts, such as the age of enrollment (a higher age corresponds to a higher probability of dropping out), type of high school degree, and tax exemption. Furthermore, high school scores are highly significant (students with higher proficiency during high school have a lower probability of dropping out).

The departments show different behaviors in terms of dropout probabilities. Finally, we can see that participation in at least one orientation event has a protective influence against dropouts.

We can also compare these results with the variable importance of the AdaBoost model, as reported in Table 10.

We observe that the selected variables for the AdaBoost model are also significant in the logistic model. The findings are consistent across the models, indicating a high degree of reliability and providing favorable evidence for the results' validity.

Table 10. Variable Importance of AdaBoost model, average values between the 10 cross-validated models (Model including orientation events).

Variable	Importance (AdaBoost)
University Department	31.44
High school score	20.72
Age enrollment	11.70
Type of High School	9.58
Area	8.21
Tax exemption	4.45
Difference in age of enrollment and high school degree	4.60
Type of students to the course	2.48
Events	2.48
Gender	2.14
Type of Course	1.30

Table 11 shows the stratification across the different levels of risk in the Department of Law.

Table 11. Distribution of observed dropouts and non-dropouts in the three levels of risk defined by the model, with marginal percentages.

	Low risk	Middle risk	High risk	Total
No drop out	52	77	40	169 (77%)
Drop out	9	25	52	86 (23 %)
Total	61 (24%)	102 (40%)	92 (36%)	255

3.3 Department analysis

The results obtained by the predictive models can be deployed to specify the analysis in a single department. The results obtained are important for organizing specific retention activities.

In addition to the descriptive analysis described in Subsection 3.1, the output of the predictive models is the probability of dropout for each student. The AdaBoost model returns risk profiles for different groups of students.

We grouped students into three categories of dropout: the students with predicted probability under the first quartile are considered low risk, the students above the third quartile are defined as high risk, and the others are in the medium risk area.

The final aim of this work is to identify high-risk students for each department and develop effective actions and strategies to reduce dropouts.

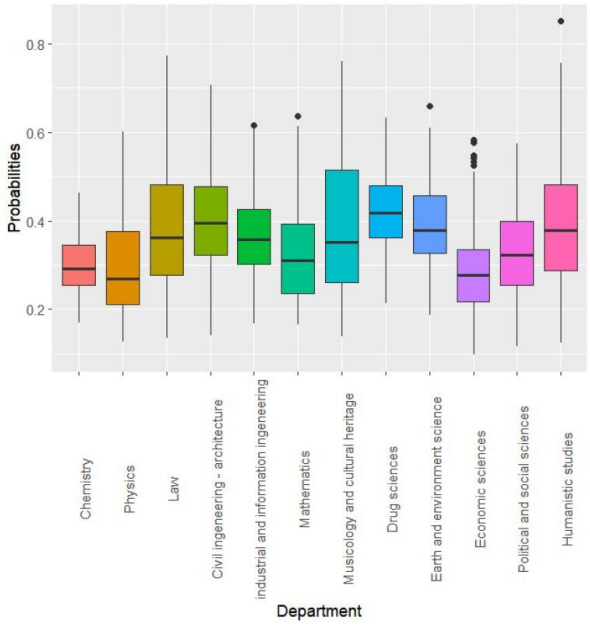


Figure 2. Distribution of dropout probability predicted by AdaBoost model in each department.

Figure 2 available in the monitoring dashboard depicts the distribution of probabilities of dropout for each department, and Figure 3 shows each department's stratification across different levels of risk (high, medium, and low).

For each department, the model produces results regarding the behavior of the students in terms of dropout. Figure 4 depicts the comparison between

dropout and non-dropout students in terms of predicted probabilities based on the data recorded in the Department of Law. That department counts 255 enrolled students with a 23% of a priori probability of dropping out. In terms of prioritization and planning of activities to reduce student dropout, we remark that this stratification introduces an effective strategy to plan retention activities. More precisely, on the basis of the results reported in Table 11 on the high-risk groups, the AdaBoost model correctly predicts 60% of the observed dropouts. Thus, providing the institutions with a clear picture to organize activities for dropout reduction.

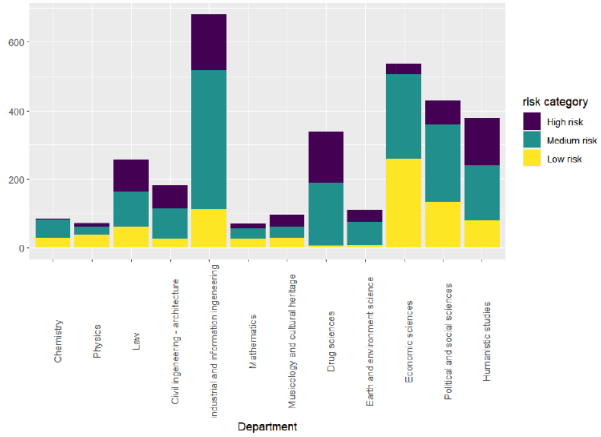
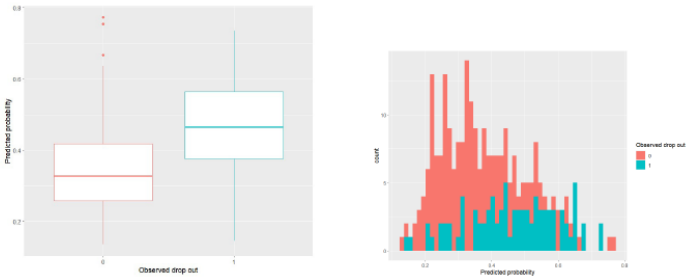


Figure 3. Stratification of students across different levels of risk in each department.



(a) Boxplot probabilities (b) Histogram probabilities

Figure 4. A Comparative Analysis of Predicted Probabilities and Observed Dropout Rates in the Department of Law: A Binary Representation of 0 for Non-dropouts and 1 for Dropouts.

Similar results are available upon request for the other departments.

4. Conclusions

The paper points out that student dropout is a more complex and multidimensional issue than most people think. The paper summarizes and discusses the main results obtained using machine learning models. Our findings demonstrate that machine learning algorithms can effectively be used in order to predict student dropout and, in particular, to distinguish significant predictors of dropout. Our contribution provides a dynamic dashboard as a practical tool for the University of Pavia to analyze recurring patterns and behaviors within a dataset of student dropouts. By analyzing the dropout rates of each area of study at the University of Pavia in relation to the number of enrollments per year, the initial assessment by the dashboard shows that areas with a high number of new enrollments tend to exhibit higher rates of student dropout. There are various factors that may contribute to this phenomenon, including but not limited to the increased diversity among students in terms of their backgrounds, socioeconomic status, and other relevant characteristics. The dataset used in our study was sourced directly from the student archives of the University of Pavia and was recorded at the time of enrollment. The results obtained from the data at hand, both in terms of methodological and computational aspects, can be extended to other university institutions, including the monitoring dashboard. Also, the inclusion of participating in orientation events, which in our study shows a negative correlation with the likelihood of dropping out, concludes that universities should focus on other non-cognitive factors such as, but not limited to, fostering students' intrinsic motivation, self-efficacy, social support, self-regulated learning strategies, and providing career counseling to enhance career decision-making and reduce the likelihood of dropouts.

References

[1] C. Barra, R. Lagravinese and R. Zotti, Does econometric methodology matter to rank universities? an analysis of italian higher education system, Socio-Economic Planning Sciences 62 (2018), 104-120. https://doi.org/10.1016/j.seps.2017.09.002

[2] M. Jadri´c, Z. Garaca and M. Cukusi´c, Student dropout analysis with application of data mining methods. Management: Journal of Contemporary Management Issues 15(1) (2010), 31-46.

[3] F. B. Rinaldi, C. G. Daniele Checchi, S. Salini and M. Turri, Ranking e valutazione: il caso delle classifiche delle universit`a. RIV Rassegna Italiana di Valutazione 41 (2009), 81-114. https://doi.org/10.3280/RIV2008-041006

[4] S. Herzog, Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. Research in Higher Education 46 (2005), 883-928. https://doi.org/10.1007/S11162-005-6933-7

[5] H. Williams and N. Roberts, i just think it' s really awkward' : transitioning to higher education and the implications for student retention 85 (2023), 1125-1141. https://doi.org/10.1007/s10734-022-00881-1

[6] J. G. Piepenburg and L. Fervers, Do students need more information to leave the beaten paths? the impact of a counseling intervention on high school students' choice of major 84 (2022), 321-341. https://doi.org/10.1007/s10734-021-00770-z

[7] N. Rotem, G. Yair and E. Shustak, Open the gates wider: affirmative action and dropping out, 81 (2021), 551-566. https://doi.org/10.1007/s10734-020-00556-9

[8] T. Zajac, F. Perales, W. Tomaszewski, N. Xiang and S. R. Zubrick, Student mental health and dropout from higher education: an analysis of australian administrative data. (2023). https://doi.org/10.1007/s10734-023-01009-9

[9] C. Bargmann, L. Thiele and S. Kauffeld, Motivation matters: predicting students career decidedness and intention to drop out after the first year in higher education 83 (2022), 845-861. https://doi.org/10.1007/s10734-021-00707-6

[10] K. Cidlinska, B. Nyklova, K. Machovcova, J. Mudrak and K. Zabrodska, Why I don't want to be an academic anymore? when academic identity contributes to academic career attrition (2023), 141-156. https://doi.org/10.1007/s10734-022-00826-8

[11] E. Ripamonti and S. Barberis, The effect of cultural capital on high school dropout: An investigation in the italian provinces, Social Indicators Research 139 (2018), 1257-1279. https://doi.org/10.1007/s11205-017-1754-6

[12] S. Meggiolaro, A. Giraldo and R. Clerici, A multilevel competing risks model for analysis of university students' careers in italy. Studies in Higher Education 42(7) (2017), 1259-1274. https://doi.org/10.1080/03075079.2015.1087995

[13] M. Cannistr`a, C. Masci, F. Ieva, T. Agasisti and A. M. Paganoni, Earlypredicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques, Studies in Higher Education 47(9) (2022), 1935-1956. https://doi.org/10.1080/03075079.2021.2018415

[14] F. Agrusti, M. Mezzini and G. Bonavolont`a, Deep learning approach for predicting university dropout: a case study at roma tre university. Journal of E-Learning and Knowledge Society 16(1) (2020), 44-54. https://doi.org/10.20368/1971-8829/1135192

[15] C. Marquez-Vera, C. R. Morales and S. V. Soto, Predicting school failure and dropout by using data mining techniques, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 8(1) (2013), 7-14. https://doi.org/10.1109/RITA.2013.2244695

[16] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems 49(4) (2010), 498-506. https://doi.org/10.1016/j.dss.2010.06.003

[17] D. Thammasiri, D. Delen, P. Meesad and N. Kasap, A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications 41(2) (2014), 321-330. https://doi.org/10.1016/j.eswa.2013.07.046

[18] D. Rodr´ıguez-G´omez, M. Feixas, J. Gairin and J. L. Munoz, Understanding catalan university dropout from a comparative approach. Procedia - Social and Behavioral Sciences 46 (2012), 1424-1429. https://doi.org/10.1016/j.sbspro.2012.05.314

[19] A. Sarra, L. Fontanella and S. Di Zio, Identifying students at risk of academic failure within the educational data mining framework, Social Indicators Research 146 (2019), 41-60. https://doi.org/10.1007/s11205-018-1901-8

[20] F. Del Bonifro, M. Gabbrielli, G. Lisanti and S. P. Zingaro, Student dropout prediction. In: International Conference on Artificial Intelligence in Education (2020), 129-140. https://doi.org/10.1007/978-3-030-52237-7

[21] M. Kadar, J. Sarraipa, J. C. Guevara and E. G. Restrepo, An integrated approach for fighting dropout and enhancing students' satisfaction in higher education. In: Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, pp. (2018), 240-247. https://doi.org/10.1145/3218585.3218667

[22] P. Perchinunno, M. Bilancia and D. Vitale, A statistical analysis of factors affecting higher education dropouts. Social Indicators Research 156 (2021), 341-362. https://doi.org/10.1007/s11205-019-02249-y

[23] D. Heredia, Y. Amaya and E. Barrientos, Student dropout predictive model using data mining techniques. IEEE Latin America Transactions 13(9) (2015), 3127-3134. https://doi.org/10.1109/TLA.2015.7350068

[24] J. Liang, J. Yang, Y. Wu, C. Li and L. Zheng, Big data application in education: Dropout prediction in edx moocs. 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), (2016), 440-443 https://doi.org/10.1109/BigMM.2016.70

[25] L. Paura and I. Arhipova, Cause analysis of students' dropout rate in higher education study program. Procedia - Social and Behavioral Sciences 109 (2014), 1282-1286. https://doi.org/10.1016/j.sbspro.2013.12.625

[26] S. Sivakumar, S. Venkataraman and R. Selvaraj, Predictive modeling of student dropout indicators in educational data mining using improved decision tree, Indian journal of Science and Technology 9(4)(2016), 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

[27] Y. Chen, A. Johri and H. Rangwala, Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. Proceedings of the 8th International Conference on Learning Analytics and Knowledge (2018), 270-279. https://doi.org/10.1145/3170358.3170410