Economic Modelling 130 (2024) 106583

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

Predicting dropout from higher education: Evidence from Italy ☆

Marco Delogu [a,b,*], Raffaele Lagravinese [c,1], Dimitri Paolini [a,d,2], Giuliano Resce [e,3]

[a] DISEA and CRENoS, University of Sassari, Italy
[b] DEM, University of Luxembourg, Luxembourg
[c] Department of Economics and Finance, University of Bari "A. Moro", Italy
[d] CORE, Catholic University of Louvain, Belgium
[e] Department of Economics, University of Molise, Italy

ARTICLE INFO

JEL classification:
C53
C55
I20

Keywords:
Early warning system
Machine learning
Dropout
Italy

ABSTRACT

Predicting university dropout is crucial. Identifying at-risk students can inform dropout prevention policies, safeguarding the nation's resources and mitigating the long-term deterioration of human capital. In contrast to previous literature, this study prioritizes predicting student dropout rather than delving into causal mechanisms. This study leverages administrative data encompassing the entire population of Italian students enrolled in bachelor's degree programs for the academic year 2013–2014. Our quantitative findings indicate that machine learning algorithms exhibit significant predictive capabilities, specifically random forest and gradient boosting machines, underscoring their potential as early warning indicators. Feature importance analysis emphasizes the role of students' first-year academic performance in dropout prediction. Furthermore, our findings provide additional evidence regarding the influence of family income, high school grades, and high school type. The adoption of these novel predictive tools can facilitate the targeted implementation of policies aimed at mitigating this phenomenon.

1. Introduction

The problem of dropout in higher education refers to the phenomenon in which students who initially enroll in a college or university fail to complete their degree programs and leave before attaining their desired qualifications. The literature has extensively analyzed the determinants of university dropout by leveraging increasingly sophisticated empirical tools along with data with ever higher levels of information. We investigate whether machine learning (ML) methods are valuable tools for predicting university dropout and find that such methods are potent predictors and can possibly be used as early warning indicators.

The issue of university dropout deserves attention for several reasons, as investments in human capital yield various benefits (Becker, 1962, 1994). First, a higher level of human capital enhances individuals' productivity, making them more valuable to employers and increasing their earnings potential.[4] Second, education improves individuals' adaptability to changing economic conditions, allowing them to acquire new skills and knowledge.

Considering that the most desired jobs require an increasingly advanced level of qualification, it is paramount to increase the share of educated individuals in the workforce. Technological change favors an educated workforce (skilling technological change) (Acemoglu, 2002; Goldin and Katz, 2008), and generates benefits in terms of economic

☆ We thank the participants to the AEDE 2022 (Porto, Portugal), ERSA 2022 (Pecs, Hungary) conferences, the economic seminar at University CY Cergy (Paris, France). We are particularly grateful to Emanuela Marroccu, the editor, and two anonymous referees for their insightful comments. The authors gratefully acknowledge financial support from: Regione Autonoma della Sardegna, Italy (Legge n. 7), DISEA, Dipartimento di Eccellenza 2018-22, University of Sassari (fondo di Ateneo per la Ricerca 2020), Italy, Fondazione di Sardegna (Economia e Benessere tra Società e Istituzioni), Italy, grant 2022-2023. We are grateful to ANVUR for sharing the data with us.

* Correspondence to: DISEA, University of Sassari, Via Muroni 23 07100, Sassari, Italy.
E-mail addresses: mdelogu@uniss.it (M. Delogu), raffaele.lagravinese@uniba.it (R. Lagravinese), dpaolini@uniss.it (D. Paolini), giuliano.resce@unimol.it (G. Resce).
[1] DEF, University of Bari "Aldo Moro" Largo Abbazia Santa Scolastica, 70124 - Bari, Italy.
[2] DISEA, Università degli studi di Sassari, Via Muroni 23, 07100, Sassari, Italy.
[3] Department of Economics, University of Molise, Via F. de Sanctis - 86100 Campobasso, Italy

[4] Psacharopoulos and Patrinos (2018) found that the private average global return for a year of schooling is 9%, which is slightly lower than the 10% estimate reported in Card (2001).

https://doi.org/10.1016/j.econmod.2023.106583
Received 5 January 2023; Received in revised form 27 October 2023; Accepted 27 October 2023
Available online 31 October 2023

growth, thus leading to improved efficiency. However, the adoption of more efficient technology does not necessarily result in a highly educated workforce. For example, a survey conducted by the ECB on leading Eurozone companies, focused on digitalization, confirmed that *"recruitment and retention of high-skill ITC staff"* is among the main obstacles to the adoption of digital technologies (see ECB Economic Bulletin Issue 7/2018.). Therefore, the relationship between supply and demand for skills can be puzzling. For instance, Carroni et al. (2023) relate the efficiency of a new technology with the effort required for learning, demonstrating that higher levels of firms' market power can impede technology adoption and individuals' education investments. Interestingly, Beladi et al. (2011) consider a model where human capital accumulation requires both capital and unskilled labor. They show that even if the growth of capital and supply of skilled labor match, issues in the unskilled intensive sector will magnify the gap between growth in demand and supply of skill. This situation is further exacerbated when individuals fail to complete their education. Finally, failure to complete higher education not only represents a waste of time and resources for students and their families, but also constitutes a misallocation of public funding, given that education is typically heavily subzidized. Despite these benefits, the percentage of students who drop out of university courses remains significantly high in many developed countries, calling for the implementation of novel policies.

Among the OECD countries, Italy is undoubtedly an emblematic case with more worrying numbers than other developed countries. As highlighted in the OECD report (OECD, 2019), although there have been some improvements in recent years, Italy still has one of the highest dropout rates among university students. Despite numerous reforms implemented over the years to increase the number of graduates, university dropout has remained a prevalent phenomenon within the Italian system (Bratti et al., 2008; Brunori et al., 2012; Oppedisano, 2011). This study contributes to the literature by employing machine learning (ML) methods to develop early warning systems that predict students at risk of university dropout. By identifying at-risk students, universities can implement proactive policies that effectively prevent student dropout, ultimately expanding the pool of graduates.

To investigate the determinants of dropout, previous work has employed standard econometric models (i.e., OLS, GLM, probit, logit, panel).[5] However, there is now consensus that these tools are intrinsically not predictive, with many authors suggesting the use of ML methods (see, e.g., Einav and Levin, 2014 and Kleinberg et al., 2015).[6]

The availability of administrative data and the increased computational power make the use of ML algorithms practical for identifying the students most at risk of dropout outlining the leading causes of it and consequently implementing targeted policies to remedy high dropout rates. If ML shows a strong ability to predict drop-out behavior such methods could eventually be used to create an early warning system that can help policymakers identify students at risk and consequently implement targeted policies.

Jia and Maloney (2015) were among the first to use econometric methods for predicting university dropout. The authors utilized administrative data collected from a university in New Zealand to assess their predictive risk model in identifying students at risk of dropout. The initial studies on the application of ML algorithms in predicting dropout primarily focused on the United States. In their work (Aulck et al., 2016) analyzed the factors contributing to first-year dropout among students at the University of Washington. Using logistic regression, random forest, and k-nearest neighbors, the authors found that grade point average scores (GPA) in math, English, chemistry, and psychology classes were the strongest predictors of student retention. Another study conducted in the United States by Sansone (2019), develops a microeconomic model where benevolent policymakers may benefit from the use of ML methods to reduce dropout rates identifying at-risk students, universities can implement proactive policies that effectively prevent student dropout, ultimately expanding the pool of graduates. Kemper et al. (2020) performed two ML approaches, logistic regressions and decision trees, to predict student dropout at the Karlsruhe Institute of Technology (KIT) in Germany. They found the most relevant single factor for predicting dropout to be combined features such as the count and the average of passed and failed examinations or average grades. Von Hippel and Hofflinger (2021) tested ML at eight Chilean universities and identified financial aid to be the main predictor of university dropout. Regarding Italy, Cannistrà et al. (2021) applied ML algorithms at the Polytechnic of Milan. Their study identified previous and early academic performance as the primary predictors of dropout. More recently, Lema et al. (2023) exploits data from two universities (Politecnico di Milano (Italy) and Vrije Universiteit Amsterdam (the Netherlands) showing that the predictive capacity of models is exchangeable.

Our work fits into this strand of the literature and enriches it. To the best of our knowledge, previous work has considered only single universities or compared a few universities with each other. In contrast, we use the Anagrafe Nazionale Studenti (ANS), a dataset produced by the Ministry of University and Research (MUR). Accordingly, we make predictions considering the whole sample, differentiating by area of study (Health, Science, Social Science and, Humanities), and by macro-region of study (North and South of Italy). The ANS collected information on all students enrolled in the Italian university system for the 2013–14 academic year. The availability of the entire population allows us to define drop-out behavior as the individual's decision to leave higher education studies; thus, we can distinguish students' decision to drop out from their choice to switch course/university. Specifically, we focus on information on undergraduate (i.e., bachelor's degree) students by following each student from enrollment to graduation or dropout by 2018, with several information items on students' academic careers and educational backgrounds. The analysis exploits a final sample of 144.904 students for whom the relevant information was available. Our main finding is that ML algorithms are potent predictors indicating their use as early warning indicators. This finding is robust to a battery of checks. In particular, we considered a battery of algorithms. Specifically, we use (1) the least absolute shrinkage and selection operator (LASSO), (2) random forest (RF), (3) gradient boosting machines (GBM), and (4) neural network (NN). The ML algorithms showing the highest predictive power were GBM and RF, with RF performing slightly better than GBM. Second, in line with Cannistrà et al. (2021), our results show that the number of ECTS (ECTS stands for European Credit Transfer and Accumulation System) earned in the first year is one of the main predictors for drop-out behavior. As detailed in Section 2, this finding is particularly relevant in light of the Italian institutional setting. It is noteworthy that to graduate in Italy, students are required to achieve a positive grade in all the exams included in their study plan. By contrast, in many other European countries, students can proceed to the next year of study as long as they obtain a positive average across the entire set of exams, allowing for the possibility of failing individual exams. Finally, our findings provide additional evidence for the role of family income, high-school grade, and high-school type. Also, RF found that the distance of the place of origin to the nearest university is an important predictor for drop-out behavior, confirming the findings of Atzeni et al. (2022).

The remainder of this paper is organized as follows. The following section describes the institutional setting of the Italian university system; Section 3 describes the dataset and features investigated in the analysis, Section 4 describes the ML models used in the study, Section 5 presents the results, and Section 6 reports the conclusions with policy suggestions derived from the results obtained.

[5] See, for Italy, Aina (2013), Belloc et al. (2010), Di Pietro (2004), Di Pietro and Cutillo (2008), Ghignoni (2017), Modena et al. (2020).

[6] Nowadays, scholars are taking advantage of ML procedures to characterize public policies through predictions (Antulov-Fantulin et al., 2021; Carrieri et al., 2021; Kleinberg et al., 2018; Mullainathan and Spiess, 2017; Athey and Imbens, 2019; Li et al., 2023; Hughes et al., 2022; Qiu and Zheng, 2023).

2. Institutional setting

This section highlights some peculiarities of the Italian university system in light of the objective of this study. The Eurydice network,[7] produces detailed information about the Italian University system relative to those of other European countries.[8]

Italy's Ministry of Universities and Research classifies university studies into *laurea classes*. Italian universities in 2013–2014 offered three types of degree: *laurea triennale*, equivalent to a bachelor's degree; *laurea specialistica*, equivalent to a 2-year master's degree; and *laurea a ciclo unico*, which combines bachelor's and master's degrees (5-year program, except for medical studies, which require a 6-year program). A class group contains courses sharing both objectives and core activities. In 2013–2014, Italian universities offered 708 different courses (degrees), belonging to 46 different classes. The ministry additionally clusters classes into four more general subject areas: (1) health; (2) science; (3) social science and (4) humanities.[9] In Italy, it is not only universities that provide first-cycle degrees; high-level arts and music education (AFAM), and higher technical institutes (politecnici) also provide similar first-cycle programs.[10] It is important to highlight that, as with other European countries, bachelor's programs provided by Italian academic institutions do not include studies across several disciplines. Among such programs are: medicine and surgery, pharmacy, veterinary science, dentistry studies, law, and architecture.[11]

In Italy, universities can be either private or public institutions. Despite their private/public status, universities act as autonomous bodies adopting their own statutes and enjoying a significant degree of freedom in terms of regulations. Given this freedom, there can be sizable differences even among public institutions. Restricting our attention solely to public institutions, some general standard lines regarding the progression of academic studies exist. Such standard practices deserve some attention in light of this study's purpose. According to the Eurydice report, students can only enroll in courses foreseen for the subsequent academic year after they have successfully completed the scheduled exams.[12] However, this statement does not hold in practice. It is most common for Italian universities to allow enrollment to the following year's courses even if a student has not passed all exams. For instance, we found that among the subset of students who graduated in time, slightly less than the 25% earned fewer than 30 ECTS at the end of their first academic year. Completing all exams requires students to earn 60 ECTS or slightly fewer. To obtain a first cycle degree, the student must earn 180 ECTS which usually includes discussing a final short essay in front of a commission. Differently than other countries, such as the UK, in Italy it is compulsory to obtain a positive grade for each course in a study plan. Also, Italian universities have some key peculiarities concerning students' evaluations for specific exams. How they conduct examinations differs in two main ways from the usual European approach. First, exams must be held at the end of the first and second semesters and after the summer break. Importantly, during each session, the same course usually has multiple examinations, with an average of six attempts per year at public universities.[13]

In Italian public universities, it is important to note that failing an exam does not hinder students from making subsequent attempts. Additionally, what may be particularly intriguing for readers unfamiliar with the Italian higher education system is that students who have received a passing grade but are dissatisfied with their results have the option to retake the exams. We point out that retaking exams after obtaining a positive grade is very common in Italy and they may decide to do so for two reasons. First, there is no official document reporting the number of attempts. Second, by retaking an exam, a student can increase her/his grades, thereby raising her/his average, which is the critical determinant for the student's final grade.[14]

3. Dataset

Our data comes from ANS national registry of students enrolled in higher education institutions in Italy.[15] We have exploited this data to implement several prediction procedures to identify risky of incurring in drop-out behavior. Remarkably, our information referred to all students who enrolled in the Italian university system, and we used three years of data on undergraduate (i.e., bachelor's degree) students who enrolled in the 2013–14 academic year. For this cohort of bachelor's degree students, we followed their academic career until the 21st of March 2018.

Our analysis excludes students enrolled in either *"laurea specialistica"* or *"laurea a ciclo unico"* university degrees, for two main reasons. First, for students enrolled in postgraduate courses, we could not retrieve information about a key variable that evidently influences drop-out behavior: the final grade obtained in the first cycle program. Also, one may argue that for individuals enrolled in postgraduate courses, the decision to drop out has different determinants than it would for students in first cycle courses.[16]

Moreover, we excluded international students, as they are selected from a different population compared to national students and constitute a self-selected group so the drop-out mechanisms for them would probably be different from those that characterize domestic students. Finally, we excluded students enrolled in online universities.[17] After applying these exclusions our dataset consisted of information on a total of 230,336 students.[18]

The next step is to differentiate between dropouts and non-dropouts. It is important to highlight that, unlike Johnes and McNabb (2004), the unique nature of the Italian university system does not allow for a clear

[7] Eurydice is a network of 40 national units based in the 37 countries of the Erasmus+ program. The network's task is to explain how education systems are organized in Europe and how they work.

[8] We collected information from several Eurydice reports; we refer interested readers to https://eacea.ec.europa.eu/national-policies/eurydice/about_en.

[9] Science was the area with most students, representing 38.4% of the sample; slightly more than the majority of students were enrolled either in humanities or social science.

[10] AFAM institutions have some crucial differences compared to universities and politecnici. In our analysis, we have not included students enrolled at AFAM institutions.

[11] These studies are organized in single-cycle courses of 5–6 years, corresponding to 300–360 CF; usually they result in the higher level, single-cycle, *laurea magistrale*.

[12] Interested readers should refer https://eacea.ec.europa.eu/national-policies/eurydice/content/second-cycle-programmes-39_pl. The document reports that *"students who do not pass the scheduled exams cannot attend courses foreseen for the following academic year"*.

[13] Also, it is common for university courses to conduct examinations in the middle of the semester, and quite often lecturers even allow additional exams during the academic year.

[14] Furthermore, an additional rationale behind this behavior stems from the significance of final grades in Italy, particularly when competing for public administration positions.

[15] ANS stands for *Anagrafe Nazionale degli Studenti* or National Registry of Students. This dataset was compiled by the Ministero dell'Università e della Ricerca (MUR), (Ministry of Universities and Research).

[16] Students in second-cycle courses should be more sensitive to labor market conditions.

[17] Note that in 2013-14, online universities accounted for only 4.53% of the total population of students enrolled in bachelor's degree courses. Most students enrolled in Italian online universities are workers; therefore, their determinants for dropping out of graduate studies are likely different from those of students enrolled in other first-cycle programs.

[18] In Appendix A.1, Table A.1 reports the complete list of universities along with some key information: the Italian region where the university is located, its legal status and the average dropout rate observed.

Table 1
Definition of drop-out variable.

Student outcome	Number	Percentage
Enrolled but degree not yet obtained ($DO_i = 0$)	71.395	31,00
Changed course/university ($DO_i = 0$)	41.009	17,80
Degree obtained on time ($DO_i = 0$)	88.221	38,30
Left higher education ($DO_i = 1$)	29.707	12,90

distinction between voluntary and involuntary university dropout.[19] To define drop-out behavior we proceeded as follows. First, we classified students into four main categories: (1) students who successfully completed their degree by the 21 March 2018; (2) students who were still enrolled during the second academic year; (3) students who changed course/university the year after their first year of enrollment[20]; (4) students who left the Italian university system before the beginning of the second academic year. Only the students belonging to the fourth category were considered to be dropouts. Accordingly, we built a dummy variable DO_i that takes a value of one if a student drops out or zero otherwise. Notice that our definition of the drop-out variable aims to capture early drop-out behavior. This aligns with our objective of assessing the effectiveness of machine learning methods in determining students at risk of prematurely abandoning their academic pursuits.[21] We found that 38.30% of the students had completed their degree by the 21 March 2018, 17.8% had changed course/university, 31.3% were still enrolled without completing their studies and 12.9% had left the university system. The latter group is the one for which the dummy variable takes a value equal to one, namely the dropouts. Notice that data availability allows us to consider dropouts to be only the students who leave the pursuit of higher education, not the ones who simply change course/university and continue their higher education journey. Table 1 reports absolute numbers and percentages for each of the different categories.

Our data show a significant difference in the percentage of dropouts across the areas of study. While dropouts are equal to only 5.3% in the health/medical area, they reach the sizeable figure of 15.1% in humanities (for the other areas, we have 12.0% of students dropping out in science and 14.4% in social sciences).

In this paper, we aim to take advantage of administrative data to predict drop-out behavior. As detailed in Section 4 we use state-of-the-art ML methodology, and the *features* choice is driven by both the availability of data and the existing literature. Unfortunately, complete information for all features is not available for all students. Our final dataset contains information about 144.904 individuals, representing 63% of the population of students.[22]

The existing literature provides evidence that the characteristics of universities, the field of study, and social and economic conditions of the students' home districts are correlated with dropout rates, see Aina et al. (2022).

In the set of predictors, we include variables capturing students' demographic information.[23] We include the variable Sex, which takes the value of one if the student is classified as female or zero otherwise.[24]

Following the recent literature (see Aulck et al., 2016; Cannistrà et al., 2021; Kemper et al., 2020) that exploits ML methods, we include the number of ECTS earned in the first year among the set of features employed to predict drop-out choice.

By computing basic descriptive statistics we find that dropout rates are much higher for students from vocational high schools, and this finding holds for all areas (health, science, social science, and humanities). Accordingly, among the set of features, we include the variable HT_i, which takes a value of one if the student has earned a high school degree in a *liceo* or zero otherwise. Another important feature included is the high-school grade. A student enrolled in an Italian high school needs to achieve a minimum final grade of 60/100 in order to graduate.[25] We rescale grades and define HG_i as a discrete variable that takes values in the interval $[0, 41]$.

Two additional features take into consideration the age of the student at the time of enrollment. Late enrollment in Italy can be due either to grade repetition in high school or general late enrollment. Accordingly, we include the variable $AGE_i = -1\,(Year\,of\,birth - 1995)$. Moreover, it is worth noting that certain students may complete high school at the age of eighteen if they entered primary school earlier. To account for this, we introduce a binary variable, denoted as Ant_i, which takes a value of one if the student's age was below nineteen and zero otherwise.

Another important determinant of dropout is household income, (see Checchi, 2000). In Italy, tuition fees depend on several factors, such as household income, field of study, and year of enrollment.[26] Private universities enjoy a much larger degree of freedom when setting tuition fees. Accordingly, we include the variable $Tax_{i,j,c,2013}$. As an additional proxy to family income, we include the variable $Income_o$ which is the gross average income in the student's municipality.

Following the classification outlined in Section 2 for each area (health, science, social science, and humanities), we include a dummy variable that takes a value equal to one if the degree belongs to the subject area or zero otherwise. We include the variable PP_j, which takes a value of one when the university j is private or zero otherwise. In the set of features, the variable $Size_j$ captures the size of the university and is equal to the number of first-cycle degree students enrolled at university j. Additionally, we incorporate the variable $SizeCourse_{j,c}$ which is equal to the sum of students enrolled at course c provided by university j.

Recently, Atzeni et al. (2022) reported evidence that drop-out behavior can be affected depending on whether the student enrolls at a local university or leaves the family residence to pursue higher education. We include two continuous variables, $TD_{o,d}$ and $TT_{o,d}$, to capture students' off-site status.

[19] Involuntary dropout refers to students who do not pursue their higher education journey because they have not attained the passes required to progress to the following year. As illustrated in Section 2 involuntary dropout is almost impossible to define in the Italian System.

[20] Changing course is not a feature of the study plan. Within this group of students, 33.15% opted for a course offered by a different university and department, while 32.7% stayed within the same university but switched to a different department. Additionally, 0.96% enrolled in a different university but pursued a similar course. The remaining 33.19% of students remained in the same university selecting a course offered by the same department.

[21] In Appendix C we employ an alternative definition of dropout that encompasses drop-out behavior in subsequent years, taking into account the actions of course changers. It is noteworthy that our primary findings remain consistent even when using this broader definition.

[22] In Appendix B we show descriptive statistics of our variable of interest for the restricted sample (see Table A.3, showing that there are only slight differences with the ones reported in Table 1 thus reducing the concern of selection. In the same Section, Table A.4 reports descriptive statistics for each feature considered in the prediction analysis. Finally, we conduct an inverse probability weighting analysis (see Appendix D), to evaluate the representativeness of our final sample.

[23] The information regarding the features pertains to the year when students enrolled.

[24] In our dataset, 54.2% of students are female. We find that the percentage of women who leave the university (11.2%), is lower than that of men (14.8%). This finding holds for all four areas of study (health, science, social science, and humanities). For instance, although women are under-represented in the area of science, the percentage of men who drop out is substantially larger than that of women.

[25] Students may get a mention. In this case, the grade is coded as 101.

[26] Tuition fees charged by Italian public universities are not uniformly determined by the central government. Tuition fees for Italian public universities depend on many determinants, and in particular on the student's family income and on the year of enrollment, see Beine et al. (2020).

Table 2
Confusion matrix.

		Data	
		Dropouts	Non-dropouts
Predicted	Dropouts	TP	**FP**
	Non-dropouts	**FN**	TN

TP stands for true positive, FP for false positive, FN for false negative and TN for true negatives.

Another variable that may correlate with drop-out behavior, (see Card, 1993), is the distance from the student's place of residence to the nearest university. We include the variable $Closeness$, which takes a value of one if the student enrolls at a university located in a different district than his/her place of residence or zero otherwise. Table A.2 in the Appendix, provides brief details of definitions, data sources, and remarks about all the features employed in our analysis.

4. Methods

Each student i has an associated target binary variable DO_i (dropout) that takes a value of one (positive sample) if the student does not complete the academic carrier, or a value of zero (negative sample) otherwise. Based on the set of features $\left(Features_{x(i)}\right)$ for student i, the prediction task is to find the function $f(.)$ (machine learning model)predicting dropout $\left(DO_i\right)$:

$$\left\{Features^{t}_{x(i)}\right\} \xrightarrow{f(.)} DO_i. \tag{1}$$

The standard approach in the ML literature is to randomly divide the data in a training set, in which the model is built and tuned, and a testing set, on which its predictive power is tested (Antulov-Fantulin et al., 2021; Cerqua et al., 2021a). The size of these two sets must be chosen while taking into account the trade-off between the benefits of a large training set (i.e., it is the only part of the database in which the algorithm builds the mapping) and the benefits of a quite large testing set (in order to reduce the testing error). Spending too much on training ($> 80\%$)) will not enable a good assessment of predictive performance because it may find a model that fits the training data very well will do poorly in the testing data (overfitting). In contrast, too much spent on testing ($> 40\%$) will not enable a good assessment of model parameters (Boehmke and Greenwell, 2019). To account for this trade-off, we follow one of the most common procedures in the literature, which is to randomly divide the database with the 80% of the data for training and the remaining 20% of observations left for the out-of-sample test set (Friedman et al., 2001). The hyper-parameter optimization is only done on the training set using a tenfold repeated cross-validation with five repetitions. All models have been implemented using R software trained with the optimization algorithms available through the `caret` package (Kuhn, 2021). Specifically, in our analysis we employed four different models:

- the Least Absolute Shrinkage and Selection Operator (LASSO): A regression statistical method that performs features selection and regularization, with L1 norm, aimed to reduce overfitting and to increase prediction accuracy and interpretability of the results (Tibshirani, 1996);
- the Random Forest (RF): A family of randomized tree-based classifier decision trees that uses different random subsets of the features at each split in the tree (Breiman, 2001). In short, Random Forest (RF) is an ensemble learning approach that combines multiple decision trees to enhance prediction accuracy.
- the Gradient Boosting Machines (GBM): The ensemble method that works in an iterative way whereby at each stage new learner tries to correct the pseudo-residual of its predecessors (Friedman, 2000). Stated differently, each model is trained to correct the mistakes made by the ensemble of models that came before it;
- the Neural Network (NN): The model that uses a set of connected input/output units in which each connection has a weight associated, and learns by adjusting the weights to predict the correct class label of the given inputs (Ripley et al., 2016). A NN is structured with multiple layers, encompassing an input layer, one or more hidden layers, and an output layer. Each layer comprises interconnected nodes that conduct computations on the input data.

These four models are the most common algorithms used in the economic literature.[27] Similarly, we rely on these different machine learning algorithms to assess their generalization abilities and see how well they perform on new, unseen instances. In particular, as different algorithms make different assumptions about the data and the problem at hand, by trying various models, one can match the algorithm's assumptions to the characteristics of the prediction problem.

Several metrics exist to evaluate the prediction power of machine learning methods. All these measures rely on comparing the value predicted by the model with the actual ones taking advantage of the testing data. In our binary classification problem, the dropouts belong to the positive class, while the non-dropouts constitute the negative class.

Computing the ratio of correct guesses with the total of guesses, $\frac{(TP+TN)}{(TP+TN+FP+FN)}$, gives a first measure of the prediction power of the model, namely the Accuracy. The other two widely used measures are sensitivity and specificity. Sensitivity (also known as true positive rate, TPR) is the ratio of students with high drop-out risk who are correctly categorized as high drop-out risk (true positive) and the total number of positive samples (high drop-out-risk students), which coincide with the probability that a drop-out student is correctly classified, $\frac{TP}{TP+FN}$. Conversely, the specificity is the probability that successful students are classified as such, $\frac{FP}{TN+FP}$.[28] However, a major drawback is that comparing the confusion matrix, as shown in Table 2 (which provides a tailored example), can become quite cumbersome when assessing multiple models. As a result, it is more common to evaluate the performance of the prediction procedure by analyzing the receiver operating characteristic (ROC) curve (Fawcett, 2006). The ROC curve shows the classifier's diagnostic ability by plotting the true positive rate (TPR), also known as sensitivity, on the y-axis against the false positive rate (FPR), equal to 1-specificity, on the x-axis. By doing so, we can easily compare the prediction power of several models by allowing the discrimination threshold to vary (Antulov-Fantulin et al., 2021). Fig. 1 provides an example of a ROC curve.[29]

Although ML algorithms show robust predictive power, they are often criticized for acting like black boxes; as such, they do not allow researchers to understand the process followed, by the algorithm, to produce the predictions. However, this criticism is unfair, given that ML methods also provide information on how useful each feature is in the prediction task by determining their weight. This procedure is known as feature importance and is determined differently for each method,

[27] Focusing on the Italian context, recent works have leveraged the potential of these algorithms to predict the bankruptcy of local governments (Antulov-Fantulin et al., 2021), vaccine hesitancy in municipalities (Carrieri et al., 2021), to estimate local mortality and local inequality during the COVID-19 pandemic (Cerqua et al., 2021b; Cerqua and Letta, 2022), and to predict Geographical Indications areas (Resce and Vaquero-Piñeiro, 2022).

[28] Other widely used measures to evaluate the goodness of fit of machine learning algorithms are positive predicted value negative predicted value, prevalence, detection rate, detection prevalence, and balanced accuracy.

[29] When the classification task is unpredictable, the negative class theoretical distribution over feature space coincides with the positive class theoretical distribution, implying that the ROC curve would be the diagonal line with an area under the curve (AUC) of 0.5. A perfect classifier has AUC equal to 1.0; the higher the AUC, the more predictive the model is. Fig. 1 provides an example of the ROC curve highlighting the AUC.

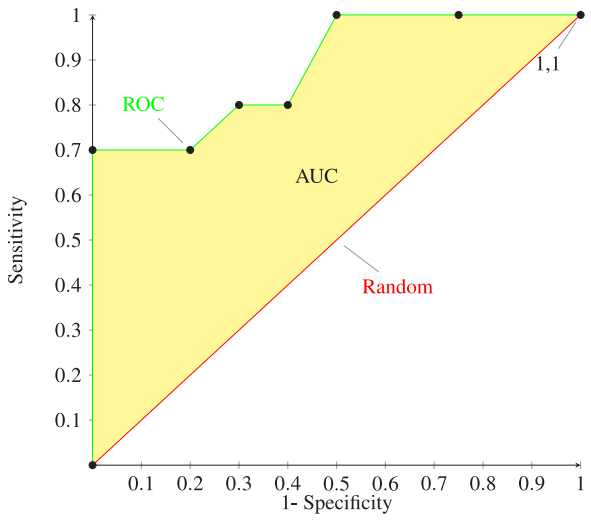

Fig. 1. Example ROC curve.
At point (1,1), the FPR equals the TPR; thus all individuals are classified as dropouts. At point (0,1), the FPR equals zero and TPR equals 1; thus, all instances are correctly classified.

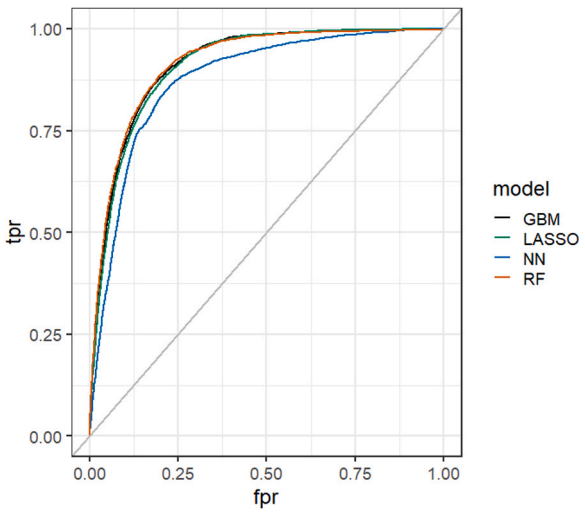

Fig. 2. ROC curve for each ML model.
ROC curve for our four models. All Models were trained on 80% of observations and tested on the remaining 20%.

(see Friedman et al., 2001). In LASSO, feature importance is estimated as the absolute value of the coefficients corresponding to the tuned model. For RF, feature importance is the mean gain produced by the feature over all the trees captured by the change in the Gini index. The feature importance in GBM is the average improvement of the splitting of the features across all the trees generated by the boosting algorithm. The feature importance in NN is determined by identifying all weighted connections between the layers in the network.

5. Results

This Section presents the results of the ML models predicting student dropout. The focus is on two main aspects: the predictability of our dependent variable (Section 5.1) and the features' importance for the independent variables used for the predictions (Section 5.2). Section 5.3 evaluates the performance of the models by providing separate analyses for northern and southern students.

5.1. Predictability of dropout

Fig. 2 shows the ROC curves for the four models (GBM, LASSO, NN, RF) trained on 80% of observations (115.924) and tested on the remaining 20% of them (28.980). The estimates are based on the cross-validation algorithm that trains and tests the model by tuning the hyper-parameters with the aim of maximizing the area under the ROC curve. The best model in terms of area under the curve (AUC) is RF (0.9155), followed by GBM (0.9128), LASSO (0.9088), and NN (0.8753) which obtains the worst performance.

Table 3 shows the four models' respective performances according to the standard measures used in the ML literature. Overall, the accuracy is statistically higher than the no-information rate for all four models used here (RF, GBM, NN, LASSO, logistic).[30] Table 3 shows that the RF and GBM models consistently outperform the other models across all metrics employed. Comparing the two best models, RF has slightly higher accuracy, and specificity, while GBM has a slightly higher sensitivity. These results, in line with previous empirical applications, confirm that the tree-based models are the best methods for structured binary tasks (Antulov-Fantulin et al., 2021; Carmona et al., 2019).

Sansone (2019) argues that the most relevant metric predicting dropout rates is the sensitivity, which captures the algorithm's ability to detect dropouts. Evidently, with a reasonably high sensitivity value, the ML algorithms could detect the students at high risk of dropout, for whom policies aimed to reduce the fraction of students leaving academic studies could be implemented. In our benchmark estimations, the sensitivity value is slightly lower than 50% in both RF and GBM, the algorithms that perform better in terms of accuracy. Although such values may seem low, our results are in line with or even better than the ones previously reported in the literature, (see Kemper et al., 2020; Sansone, 2019). As an additional empirical application, we also test the predictability of dropout within each area of study (health, science, social science, and humanities). Table 4 shows that droput rates are predictable for three out of the four areas.[31] Focusing on the algorithms that perform better in terms of accuracy, we observe that this measure is statistically higher than the no information rate for science, social science, and humanities. However, in the case of health, our measure of accuracy is not statistically higher than the no-information rate.

It is worth noting that for both social science and humanities, our best models exhibit a true positive rate (sensitivity) exceeding 50%. This indicates that the utilization of clustering techniques based on study areas enhances the predictive capability of regression tree methods.

5.2. Features importance

This section shows the most important features of the prediction task. In the previous section, we observed that the two top-performing models, namely RF and GBM, utilize combinations of diverse regression

[30] Table 3 also reports the performance of logistic regression, which performs poorly compared to other ML models, in line with the LASSO performance. This is not surprising given that the logistic regression is, by construction, the LASSO without regularization that is commonly implemented to avoid over-fitting (i.e., to improve performance on the unseen dataset).

[31] It should be noted that the health area primarily consists of students enrolled in nursing studies. In Italy, each university offering health courses can admit only a limited number of students, with placements allocated based on the results of an entrance test, for which we lack information. We believe that the low dropout rates observed in this area can be attributed mainly to the high demand of these skills in the labor market, upon completion of their studies.

Table 3
Performance of the models.

	RF	GBM	NN	LASSO	Logistic
Accuracy	0.898	0.895	0.876	0.891	0.891
95% CI	(0.8945, 0.9015)	(0.8912, 0.8983)	(0.8726, 0.8802)	(0.8876, 0.8948)	(0.8877, 0.8949)
No Information Rate	0.871	0.871	0.871	0.871	0.871
P-Value [Acc > NIR]	0.000	0.000	0.005	0.000	0.000
Sensitivity	0.471	0.443	0.183	0.386	0.389
Specificity	0.961	0.962	0.979	0.966	0.966
Pos Pred Value	0.643	0.630	0.561	0.626	0.625
Neg Pred Value	0.925	0.921	0.890	0.914	0.915
Prevalence	0.129	0.129	0.129	0.129	0.129
Detection Rate	0.061	0.057	0.024	0.050	0.050
Detection Prevalence	0.094	0.091	0.042	0.079	0.080
Balanced Accuracy	0.716	0.702	0.581	0.676	0.677
AUC	0.916	0.913	0.875	0.909	0.909

Table 4
Models' performances within each area of study.

	Health (GBM)	Science (RF)	Social Science (RF)	Humanities (GBM)
Accuracy	0.945	0.885	0.891	0.901
95% CI	(0.9287, 0.9577)	(0.8752, 0.8938)	(0.8819, 0.8996)	(0.8895, 0.9118)
No information rate	0.942	0.858	0.848	0.846
P-Value [Acc > NIR]	0.376	0.000	0.000	0.000
Sensitivity	0.1765	0.3801	0.5602	0.5404
Specificity	0.9933	0.9625	0.9555	0.9486
Pos pred value	0.6154	0.5754	0.6924	0.6605
Neg pred value	0.9522	0.9207	0.9239	0.9178
Prevalence	0.0571	0.1179	0.1518	0.1561
Detection rate	0.0101	0.0448	0.0850	0.0843
Detection prevalence	0.0164	0.0779	0.1228	0.1277
Balanced accuracy	0.5849	0.6713	0.7578	0.7445
AUC	0.906	0.899	0.929	0.924

Goodness of fit measures for the best model (among RF, GBM, NN, and LASSO) in terms of AUC for each area of study.

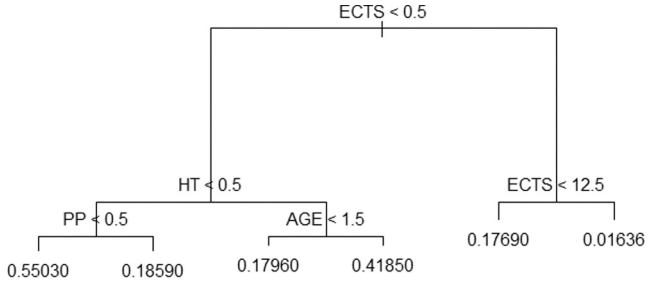

Fig. 3. Regression tree over the whole sample.
Regression tree with three branches. Most important features are: ECTS, HT and PP.

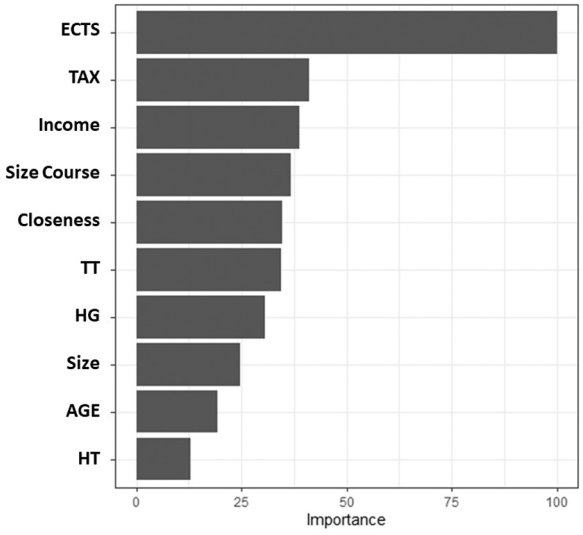

Fig. 4. Feature importance for predicting dropout: The first 10 important features in RF.
Feature Importance to predict university dropout for the first 10 important features in Random Forest. Random forest trained on 80% of observations and tested on the remaining 20%.

trees. While a standalone standard regression tree may have limited predictive power, it effectively highlights the most significant variables within the prediction task. We first consider a standard regression tree with three branches to uncover the most important features in the prediction task.

From Fig. 3 it can be observed that the number of ECTS earned by the student at the end of the academic year is the most critical factor for explaining dropout, as this feature is on the top of the tree. For students who did not earn any ECTS, the type of high school (HT(*liceo* vs. *non liceo*) and the type of university (public vs. private) PP are important factors for explaining the drop-out behavior. The likelihood of dropout substantially reduces for students who did not earn ECTS in the first year that they were enrolled in private universities and with a *liceo* diploma. Such results confirm the importance of the socioeconomic background explaining the choice to drop out from higher education. Private universities ask for more considerable tuition fees, thus fewer students from less advantaged households enroll in such universities. It is evident that for students with similar characteristics (no exam passed in the first year and with high school education at a vocational school) the likelihood of dropout is more than four times higher among those enrolled in private universities. Another, complementary, explanation of this finding is that private universities are likely to implement policies that eventually reduce their share of dropouts. From the right side of the decision tree, we see that the dropout probability is almost equal to zero for students who earned more than 12.5 ECTS during their first year.

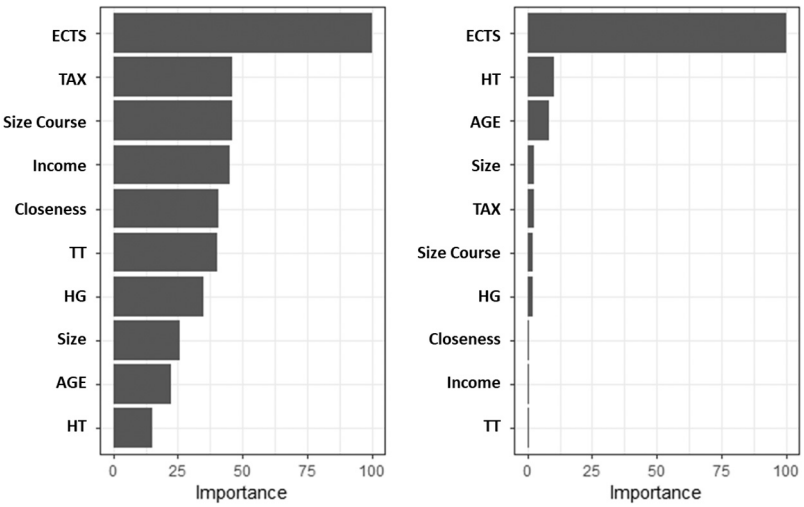

Fig. 5. Feature importance for predicting dropout by territorial area: the first 10 important features in the northern region (left panel) and southern region (right panel).
Feature Importance to predict university dropout for the first 10 important features in RF (left panel, northern region) and GBM (right panel, Southern regions). Both models were trained on 80% of observations and tested on the remaining 20%.

Table 5
Models' performance across geographical areas (northern and southern regions).

	North (RF)	South (GBM)
Accuracy	0.9080	0.8688
95% CI	(0.9039, 0.9119)	(0.8612, 0.8761)
No information rate	0.8888	0.8258
P-Value [Acc > NIR]	0.0000	0.0000
Sensitivity	0.4210	0.5273
Specificity	0.9689	0.9408
Pos pred value	0.6292	0.6528
Neg pred value	0.9304	0.9042
Prevalence	0.1113	0.1742
Detection rate	0.0468	0.0919
Detection prevalence	0.0744	0.1407
Balanced accuracy	0.6950	0.7340
AUC	0.917	0.907

Goodness of fit measures for the best models separately for northern- (RF) and southern (GBM) region.

Fig. 4 reports the first ten important features for predicting dropout in RF. Similar results, available upon request, are obtained when determining feature importance for GBM. In line with the decision tree reported in Fig. 3 the most important feature is the number of ECTS earned by the students at the end of the academic year.[32] Notice that ECTS obtained in the first year is information available to the university, which could easily detect students at risk of dropout. However, adding additional features and using the proper ML methods improves prediction power. ECTS is followed by TAX, average gross income in the municipality of origin, size of the course, distance in terms of time between the student's place of residence and the university, the physical distance between the student's place of residence and the university, the high school grade, number of first-cycle degree students enrolled at university, age, and type of the high school attended by the student. Finally, our feature importance analysis is in line with the results reported in the literature, (see Aina et al., 2022). Furthermore, in line with Atzeni et al. (2022), we see that the location choice of the students, captured by the variables TT and $Closeness$, contains information that is also valuable for predicting drop-out behavior.

Fig. 4 while showing the most important features does not provide insight on both the relationship between the features and the dropout, but only highlights the importance of these features in the prediction analysis. While we do not have the ambition to uncover causal relationship in this study, Figure Fig. A.2 of Appendix F of the appendix reports the partial plots for the three most important features (see Fig. 4): ECTS, Tax, and Income.

It should be noted that when missing observations are removed from our database, there is a potential risk of compromising the representativeness of the data. To ensure the generalizability of our findings to the entire population of Italian universities, we conducted a repeated analysis where we assigned weights to selected observations based on their representativeness. This was achieved through inverse probability weighting (IPW), see Appendix D. The results, presented in the Appendix (Appendix D Table A.7 and Fig. A.1), confirm the performance of the models shown in Table 3, as well as the importance of features in the best model (Random Forest).

In the following section, we evaluate the robustness of our result by considering separately students enrolled in southern and northern universities.

5.3. North-south heterogeneity

In this section, we assess the predictability of dropout by dividing the sample into northern students ((102.702) observations) and southern ((40,012 observations students). The **last decade** has been characterized by the profound financial crisis, exacerbating the economic differences between North and South (see about this aspect (Lagravinese, 2015)). In terms of AUC, RF is the best algorithm for predicting dropout in northern regions, while GBM is the best algorithm for predicting dropout in southern regions.[33] Dropout in northern regions has higher predictability than that in southern regions in terms of AUC, but the accuracy is statistically higher than the no-information rate in both areas (see Table 5). It is important to note that when considering southern students separately, we observe a higher value for sensitivity. Interestingly, ML methods exhibit improved performance in identifying students at risk of dropout within southern universities, where dropout

[32] Feature importance of other features is determined relative to the most important one, ECTS. We investigate whether the predictive power of the ECTS is determinant for the model performance conducting an additional where we exclude from the set of features the ECTS obtained during the first year. The results, reported in Table A.8 in Appendix E of the appendix, indicate that the prediction performance without ECTS becomes unsatisfactory, with Accuracy being not statistically larger than the No Information rate.

[33] Table 5 shows the models with the best accuracy for both sub-samples. Results obtained using the other algorithms are available upon request.

rates are typically higher. For those ML algorithms, we conduct feature importance analysis (Table 5).

From Table 5 certain degree of heterogeneity emerges in terms of feature importance. Fig. 5 shows that while ECTS is the most important feature in both geographical areas, the relative importance of ECTS in the southern region is higher than the reactive importance of ECTS in the northern regions. The ten most important features are the same in both regions, but the ranking and magnitude in relation to the ECTS change for all remaining features. In line with the results of Stinebrickner and Stinebrickner (2012), our results suggest that learning about their own ability is very important for southern students. Low results in terms of ECTS during the first year discourage them from pursuing any higher education studies. By contrast, for northern students, the higher prediction power of the TAX variable is remarkable, which likely captures students coming from less advantaged households (in line with this result, notice the high importance of the $Income$ variable). Such results suggest that northern students with low academic achievement and from poorer households are more likely to drop out of higher education. Notice that the low predictive power of the TAX variable when considering southern students separately, can be attributed in part to the availability of numerous grants specifically designed for students from economically disadvantaged households residing in southern regions.

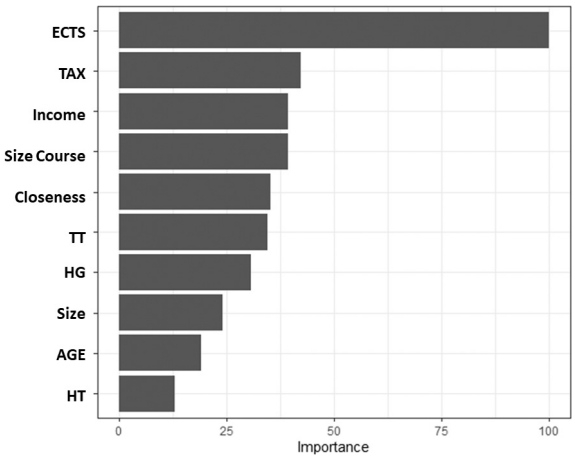

Fig. A.1. Feature importance for predicting dropout with weighted model: The first 10 important features in RF.
Feature Importance to predict university dropout for the first 10 important features in Random Forest with inverse probability weighting.

6. Conclusions

This paper investigated whether ML methods are suitable for identifying higher education students at risk of dropping out. We demonstrated how various machine learning algorithms can better and more accurately predict students who drop out of university. Unlike the rest of this novel literature, we took advantage of a dataset that comprises the entire set of Italian bachelor's students. Accordingly, we could define drop-out behavior as leaving university, instead of merely dropping out from a university course. Our paper considered a battery of ML methods, showing the best algorithms in solving the prediction task to be random forest and gradient boosting machine.

Our results showed that the strongest predictor of drop-out behavior were the few ECTS earned during the first year of study. This finding is interesting in light of the Italian institutional setting. In Italy, students need to obtain a positive grade for each exam in their study plan, whereas in most other EU countries, students are required to attempt all exams and may be allowed to enroll for the subsequent year even without obtaining a positive grade in some of them. Also, Italian students are allowed to retake exams (even after obtaining a positive grade), and the number of attempts, for each exam is usually equal to or larger than six yearly. Interestingly, our models demonstrate a higher predictive capability in Southern universities, where the issue of dropout is particularly severe and the socio-economic context is more disadvantaged compared to the central and northern regions.

Our analysis has several policy implications. First, dropout is predictable, and ML algorithms (specifically regression tree methods) can be used to identify students at risk. The use of these new predictive tools could facilitate the implementation of targeted policies to mitigate this phenomenon. In line with this argument, Bettinger and Long (2009) showed that such programs are effective in reducing the likelihood of dropout. Furthermore, the importance of the ECTS can be linked to the peculiarity of the Italian higher education system. Löfgren and Ohlsson (1999) showed that students perform worse with more relaxed rules. The possibility of retaking exams seems to be one of the possible candidates for explaining the poor results that Italian students achieve during their first year in terms of ECTS relative to the necessity for getting a positive grade in each exam in a study plan.

Addressing the drop-out problem requires a multifaceted approach. Several programs can include academic assessments, counseling, and mentoring programs to address any challenges faced by at-risk students, some of whom may struggle with the rigor of their coursework, find it difficult to adapt to academic demands or experience a mismatch between their skills and the program's requirements. These challenges can lead to frustration and a loss of motivation, prompting students to leave. Moreover, it could be useful to establish mentorship programs (especially in the first year of study) where students can connect with faculty, staff, or alumni who can provide guidance, support, and encouragement. Additionally, counseling services should be offered to address personal, emotional, and mental health issues that contribute to dropout rates.

Further research should investigate whether the peculiar characteristics of the Italian system partially explain dropout behavior and late graduation. In addition, the legal value of the final grade for selection for public offices, along with the non-consideration of the time that students take to complete their studies might incentivize them to continue their university studies at a slow pace. Given the high dropout rate in the first year, further and more effective university guidance policies should be implemented. For this purpose, the best programs for university orientation, retention/success, and satisfactory experience should be carefully investigated (Eather et al., 2022). Our feature analysis suggests that among the main factors contributing to dropout, tuition fees, and income are crucial. In Italy, a significant problem is the scarcity of accommodations for out-of-town students and high rental costs. Therefore, it is crucial to ensure that students have access to financial aid options, scholarships, and grants to alleviate the financial burden associated with education. Providing substantial financial support can significantly reduce the likelihood of students dropping out owing to financial constraints. By addressing these challenges, institutions can increase student retention and completion rates, thereby ensuring that more students obtain higher education qualifications.

Declaration of competing interest

The authors have no conflicts of interest to declare. All co-authors have seen and agreed with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other journal.

Data availability

The authors do not have permission to share data.

Appendix A

A.1. Descriptive statistics

Table A.1 reports the region, name, legal status, and dropout rate (as defined in Section 3) for each university included in the dataset.

The table below reports each feature, its definition, and data sources, along with some remarks.

Appendix B. Descriptive statistics for the restricted sample

As described in Section 3, our empirical analysis is based on a sub-sample of 144.904 individuals due to missing variables. Table A.3 indicates that the dropout statistic in this sub-sample closely resemble those of the full sample, which included complete information on all features.

Appendix C. Late dropout

In this Section, we expand our main analysis by modifying the definition of the drop-out variable. We now include among dropouts all students who left the university system after the second year of education. With this extended definition, the percentage of students who drop out rises to 27%. To explore the implications of this extended definition, we conducted a similar to the one presented in Table 3 in Section 5. The findings from this analysis are summarized in Table A.5:

The top-performing models in terms of accuracy remain RF and GBM, with the former demonstrating slightly superior performance. Intriguingly, with the adoption of this new definition, the accuracy decreases while the sensitivity improves. Therefore, using the extended definition, our algorithms do a better job of identifying the students at risk. Table Table A.6 reports the feature importance of our best model (Random forest). The results almost overlap with the ones of Fig. 4.

Appendix D. Results with inverse probability weighting

In this section, we show the results of the models weighting selected observations based on their representativeness (through inverse probability weighting - IPW). Weights are estimated based on stratifying the students' population by gender and the Italian region where the university is located.

Appendix E. Prediction of dropout without ECTS

In this section, we show the results of the models without ECTS among the predictors.

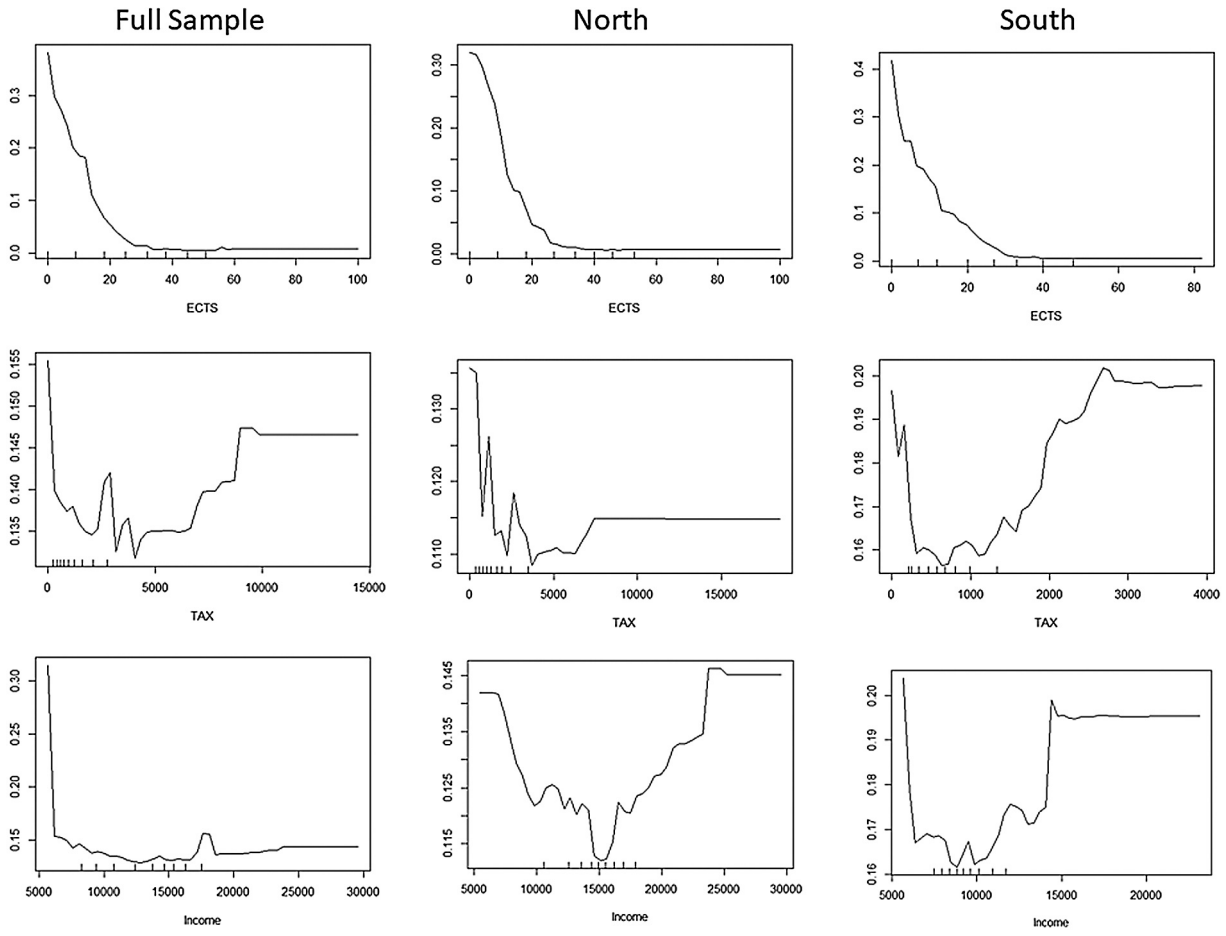

Fig. A.2. Partial Plots for ECTS, TAX and Income.

Table A.1
Academic institutions.

Region	University	Legal status	Dropout rate (%)
ABRUZZO	University G. d'Annunzio in Chieti-Pescara	public	14.6
ABRUZZO	University of L'Aquila	public	18.1
ABRUZZO	University of Teramo	public	17.8
BASILICATA	University of the Basilicata	public	17.8
CALABRIA	University Magna Graecia	public	17.0
CALABRIA	University of Calabria	public	14.9
CALABRIA	University for Foreigners Dante Alighieri	private	16.0
CAMPANIA	University of Naples Parthenope	public	24.3
CAMPANIA	Suor Orsola Benincasa University	private	16.0
CAMPANIA	University of Sannio	public	14.2
CAMPANIA	University of Salerno	public	13.7
CAMPANIA	"Orientale" University of Naples	public	15.4
CAMPANIA	Second University of Naples	public	12.8
CAMPANIA	University of Naples Federico II	public	12.8
EMILIA ROMAGNA	University of Parma	public	14.4
EMILIA ROMAGNA	University of Ferrara	public	12.6
EMILIA ROMAGNA	University of Modena and Reggio Emilia (UNIMORE)	public	14.0
EMILIA ROMAGNA	University of Bologna	public	10.5
FRIULI VENEZIA GIULIA	University of Udine	public	14.6
FRIULI VENEZIA GIULIA	University of Trieste	public	12.0
LAZIO	University of Rome Tor Vergata	public	14.4
LAZIO	University of Rome Foro Italico	public	8.9
LAZIO	Rome University of International Studies	private	10.0
LAZIO	LUISS Guido Carli	private	1.9
LAZIO	European University of Rome	private	6.2
LAZIO	Link Campus University	private	6.7
LAZIO	Campus Bio-Medico University	private	4.7
LAZIO	Roma Tre University	public	16.6
LAZIO	Free University Maria SS.Assunta (LUMSA)	private	9.7
LAZIO	University of Tuscia	public	20.0
LAZIO	Sapienza University of Rome	public	6.2
LAZIO	University of Cassino and Southern Lazio	public	19.6
LIGURIA	University of Genova	public	15.6
LOMBARDIA	University of Milano-Bicocca	public	10.3
LOMBARDIA	University of Pavia	public	10.0
LOMBARDIA	Politecnico di Milano	public	4.4
LOMBARDIA	University of Milano	public	15.0
LOMBARDIA	University of Bergamo	public	18.2
LOMBARDIA	University of Insubria	public	13.2
LOMBARDIA	Università Bocconi	private	0.8
LOMBARDIA	University of Brescia	public	11.9
LOMBARDIA	Università Cattolica del Sacro Cuore	private	7.7
LOMBARDIA	Free University of Languages and Communication	private	10.7
LOMBARDIA	LIUC – Università Cattaneo	private	7.2
LOMBARDIA	Vita-Salute San Raffaele University	private	1.8
MARCHE	University of Camerino	public	14.2
MARCHE	University of Urbino Carlo Bo	public	15.8
MARCHE	Marche Polytechnic University	public	10.8
MARCHE	University of Macerata	public	14.2
MOLISE	University of Molise	public	17.3
PIEMONTE	Politecnico di Torino	public	8.1
PIEMONTE	University of Gastronomic Sciences	private	0.0
PIEMONTE	University of Turin	public	13.1
PIEMONTE	University of Piemonte Orientale "Amedeo Avogadro"	public	13.6
PUGLIA	University of Salento	public	15.8
PUGLIA	University of Foggia	public	21.8
PUGLIA	LUM Jean Monnet University	private	16.8
PUGLIA	University of Bari Aldo Moro	public	20.0
PUGLIA	Polytechnic of Bari	public	10.6
PUGLIA	Università Mediterranea of Reggio Calabria	public	31.3
SARDEGNA	University of Sassari	public	15.7
SARDEGNA	University of Cagliari	public	16.6
SICILIA	University of Palermo	public	15.4
SICILIA	Kore University of Enna (UKE)	private	19.7
SICILIA	University of Catania	public	15.7
SICILIA	University of Messina	public	18.7
TOSCANA	University for Foreigners of Siena	public	13.8
TOSCANA	University of Pisa	public	11.2
TOSCANA	University of Siena	public	10.1
TOSCANA	University of Florence	public	14.7
TRENTINO ALTO ADIGE	University of Trento	public	9.8
TRENTINO ALTO ADIGE	Free University of Bozen-Bolzano	private	10.9
UMBRIA	University for Foreigners Perugia	public	14.0

(*continued on next page*)

Table A.1 (*continued*).

UMBRIA	University of Perugia	public	15.3
VALLE D'AOSTA	Università della Valle d'Aosta	private	14.0
VENETO	University of Verona	public	12.0
VENETO	Ca' Foscari University of Venice	public	9.2
VENETO	University IUAV of Venice	public	6.0
VENETO	University of Padova	public	10.6

Table A.2
Data sources and definitions.

Variable	Definition	Source	Remarks
Dropout variable $(DO_{\{i,\}})$	Dummy variable that takes a value of one when the student drops out from the course/university or zero otherwise.	ANS data; our computation.	seeSection 3 for details
Sex	dummy variable that takes the value one if the student is classified female and zero otherwise	ANS data.	
$ECTS_{i,2013}$	Number of ECTS earned by students during the first academic year.	ANS data.	
HT_i	Dummy variable that captures the type of high school i.	ANS data.	The variable takes a value equal to one only if the high school is a*liceo* of the traditional type, either*classico* or*scientifico*. For all the other high schools, the variable is set equal to zero.
HG_i	Variable capturing the high school grade of student i.	ANS data.	The minimum grade to obtain a high school certificate in Italy is equal to 60, the maximum is equal to 100 (however, students may obtain a mention). We scale by subtracting 60 from each vote.
AGE_i	This variable captures whether the students enrolled*late*; $AGE_i = -1\left(Yearofbirth_i - 1995\right)$.	ANS data; our computation.	Note that in Italy, students usually finish high school at the age of 19.
ANT_i	This variable takes the value one if a student enrolled at the university at an age lower than 19 and zero otherwise	ANS data; our computation.	Students who anticipated entrance to primary school may enroll in university courses at an age lower than nineteen.
$TAX_{i,j,c,2013}$	This variable captures the amount of tuition fees charged to student i enrolled to university j at course c during the academic year 2013–2014.	ANS data.	Note that in Italian public universities, tuition fees should not be paid upfront.
$Income_o$	Average gross income in the place of origin of the student.	Italian Ministry of Economics and Finance.	Raw data taken from the fiscal declaration data set available at the municipal level. Original information is split into eight classes of gross income.
$Area_{i,a}$	The subscript a captures the area of study (health, science, social science, humanities). Accordingly, we build four dummy variables.	ANS data.	The four areas of study are health, science, social science, and humanities.
PP_j	Dummy variable that takes a value of one if institution j is private or zero otherwise.	ETER dataset.	
$Size_j$	Continuous variables equal to the number of first-cycle degree students enrolled at university j.	ANS data; our computation.	
$SizeCourse_{j,c}$	Number of students enrolled at university j and first-cycle degree course c.	ANS data; our computation.	
$TD_{i,u,o}$	This variable is equal to the distance between the student's i place of residence, o and the destination university u.	ANS data; our computation.	We employed the STATA routine developed by Weber and Péclat (2017).
$TT_{i,u,o}$	This variable is equal to the distance in terms of time between the student's i place of residence, o and the destination university u.	ANS data, our computation.	We employed the STATA routine developed by Weber and Péclat (2017).
$Closeness$	Dummy variable that takes a value of one when the district in which the student enrolls hosts a university.	ANS and ETER data; our computation.	51 out of the 108 Italian districts host a university. Each Italian region hosts at least one university.

Table A.3
Definition of drop-out variable (Subsample).

Student outcome	Number	Percentage
Enrolled but degree not yet obtained ($DO_i = 0$)	44.773	30,8
Changed course/university ($DO_i = 0$)	24.746	17,1
Degree obtained on time ($DO_i = 0$)	56.768	39,2
Left higher education ($DO_i = 1$)	18.617	12,9

Table A.4
Descriptive statistics.

Feature	Obs	Mean	SD
Sex	144,904	0.543	0.498
$ECTS$	144,904	24.702	19.868
HT	144,904	0.507	0.500
HG	144,904	18.623	11.557
AGE	144,904	2.020	3.889
ANT	144,904	0.037	0.188
TAX	144,904	1,420.134	1,603.443
$Income$	144,904	13,279.600	3,530.342
PP	144,904	0.071	0.257
$Size$	144,904	5,625.661	3,359.256
$SizeCourse$	144,904	246.605	207.390
TD	144,904	1.460	2.721
TT	144,904	110.948	173.216
$Closeness$	144,904	0.502	0.500

This table reports descriptive statistics for all features considered in the analysis. For definitions and explanations of each feature see Table A.2.

Table A.5
Performance of the models.

	RF	GBM	NN	LASSO
Accuracy	0.850	0.843	0.8794	0.836
95% CI	(0.8458, 0.8541)	(0.839, 0.8474)	(0.795, 0.7988)	(0.8317, 0.8403)
No information rate	0.793	0.793	0.793	0.793
P-Value [Acc > NIR]	0.000	0.000	0.324	0.000
Sensitivity	0.561	0.533	0.032	0.539
Specificity	0.925	0.924	0.993	0.914
Pos pred value	0.662	0.647	0.545	0.619
Neg pred value	0.890	0.884	0.797	0.884
Prevalence	0.207	0.207	0.207	0.207
Detection rate	0.116	0.110	0.007	0.111
Detection prevalence	0.743	0.729	0.513	0.726
Balanced accuracy	0.883	0.877	0.810	0.870

Table A.6
Feature importance (RF).

ECTS	100
TAX	32.98
SizeCourse	30.7
Income	27.8
TD	27.8
TT	27.66
HG	24.87
Size	18.71
Age	9.64
HT	5.5

Feature importance to predict university late dropout for the first 10 important features under the best performing model (RF). The model was trained on 80% of observations and tested on the remaining 20%.

Appendix F. Partial Plots of Random Forest estimates

In this section, Fig. A.2 reports the Partial Plots obtained considering our best-performing model, Random Forest, considering the three most important features: ECTS, TAX, and Income. This analysis provides interesting insights. First, we uncover a negative association between the number of ECTS and dropout. As expected, a decrease in the number of ECTS earned increases the likelihood of dropout. For the second most important feature, TAX, the partial plot suggests a U-shaped relationship with dropout. Higher dropout rates are associated with low and high levels of TAX, while a lower dropout risk is associated with an intermediate level of TAX. Finally, we uncover a similar pattern between Income and dropout, with our figure suggesting a higher risk of dropout for students coming from municipalities with either low or high levels of income. It is worth noticing the ability of Machine learning methods to uncover this non-linear relationship. Also, we conduct a similar analysis clustering the students as in Section 5.3. It is worth noticing that these results are unaffected by the clustering of students depending on the area of origin.

Table A.7
Performance of the weighted models.

	RF	GBM	NN	LASSO
Accuracy	0.897	0.892	0.873	0.887
95% CI	(0.8937, 0.9007)	(0.8883, 0.8955)	(0.8692, 0.8769)	(0.8829, 0.8902)
No Information Rate	0.873	0.873	0.873	0.873
P-Value [Acc > NIR]	0.000	0.000	0.504	0.000
Sensitivity	0.456	0.413	0.000	0.364
Specificity	0.961	0.962	1.000	0.963
Pos Pred Value	0.632	0.610	NaN	0.586
Neg Pred Value	0.924	0.919	0.873	0.912
Prevalence	0.127	0.127	0.127	0.127
Detection Rate	0.058	0.052	0.000	0.046
Detection Prevalence	0.092	0.086	0.000	0.079
Balanced Accuracy	0.709	0.687	0.500	0.663
AUC	0.912	0.908	0.827	0.901

Table A.8
Performance of the models without ECTS.

	RF	GBM	NN	LASSO
Accuracy	0.855	0.855	0.851	0.851
95% CI	(0.8488, 0.8608)	(0.8487, 0.8607)	(0.8451, 0.8572)	(0.845, 0.8572)
No Information Rate	0.851	0.851	0.851	0.851
P-Value [Acc > NIR]	0.119	0.129	0.506	0.516
Sensitivity	0.107	0.066	0.000	0.029
Specificity	0.986	0.993	1.000	0.995
Pos Pred Value	0.565	0.608	NaN	0.496
Neg Pred Value	0.863	0.859	0.851	0.854
Prevalence	0.149	0.149	0.149	0.149
Detection Rate	0.016	0.010	0.000	0.004
Detection Prevalence	0.028	0.016	0.000	0.009
Balanced Accuracy	0.546	0.529	0.500	0.512
AUC	0.764	0.767	0.713	0.733

References

Acemoglu, D., 2002. Directed technical change. Rev. Econom. Stud. 69 (4), 781–809.

Aina, C., 2013. Parental background and university dropout in Italy. Higher Educ. 65 (4), 437–456.

Aina, C., Baici, E., Casalone, G., Pastore, F., 2022. The determinants of university dropout: A review of the socio-economic literature. Soc. Econ. Plan. Sci. 79, 101102.

Antulov-Fantulin, N., Lagravinese, R., Resce, G., 2021. Predicting Bankruptcy of Local Government: A Machine Learning Approach. J. Econ. Behav. Organ. 183, 681–699.

Athey, S., Imbens, G.W., 2019. Machine learning methods that economists should know about. Annu. Rev. Econ. 11, 685–725.

Atzeni, G., Deidda, L.G., Delogu, M., Paolini, D., 2022. Drop-out decisions in a cohort of Italian universities. In: Teaching, Research and Academic Careers: An Analysis of the Interrelations and Impacts. Springer International Publishing Cham, pp. 71–103.

Aulck, L., Velagapudi, N., Blumenstock, J., West, J., 2016. Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

Becker, G.S., 1962. Investment in human capital: A theoretical analysis. J. Polit. Econ. 70 (5), 9–49.

Becker, G.S., 1994. Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, third ed. In: NBER Books, National Bureau of Economic Research, Inc, URL https://ideas.repec.org/b/nbr/nberbk/beck94-1.html.

Beine, M., Delogu, M., Ragot, L., 2020. The role of fees in foreign education: evidence from Italy. J. Econ. Geogr. 20 (2), 571–600.

Beladi, H., Marjit, S., Weiher, K., 2011. An analysis of the demand for skill in a growing economy. Econ. Model. 28 (4), 1471–1474. http://dx.doi.org/10.1016/j.econmod.2011.02.032, URL https://www.sciencedirect.com/science/article/pii/S0264999311000502.

Belloc, F., Maruotti, A., Petrella, L., 2010. University drop-out: An Italian experience. Higher Educ. 60 (2), 127–138.

Bettinger, E.P., Long, B.T., 2009. Addressing the needs of underprepared students in higher education does college remediation work? J. Hum. Res. 44 (3), 736–771.

Boehmke, B., Greenwell, B.M., 2019. Hands-on Machine Learning with R. CRC Press.

Bratti, M., Checchi, D., De Blasio, G., 2008. Does the expansion of higher education increase the equality of educational opportunities? Evidence from Italy. Labour 22, 53–88.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Brunori, P., Peragine, V., Serlenga, L., 2012. Fairness in education: the Italian university before and after the reform. Econ. Educ. Rev. 31 (5), 764–777.

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., Paganoni, A.M., 2021. Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. Stud. Higher Educ. 1–22.

Card, D., 1993. Using Geographic Variation in College Proximity to Estimate the Return to Schooling. NBER Working Papers 4483, National Bureau of Economic Research, Inc, URL https://ideas.repec.org/p/nbr/nberwo/4483.html.

Card, D., 2001. Estimating the return to schooling: Progress on some persistent econometric problems. Econometrica 69 (5), 1127–1160.

Carmona, P., Climent, F., Momparler, A., 2019. Predicting failure in the US banking sector: An extreme gradient boosting approach. Int. Rev. Econ. Finance 61, 304–323.

Carrieri, V., Lagravinese, R., Resce, G., 2021. Predicting vaccine hesitancy from area-level indicators: A machine learning approach. Health Econ. 30 (12), 3248–3256.

Carroni, E., Delogu, M., Pulina, G., 2023. Technology adoption and specialized labor. Int. Econ. 173 (C), 249–259. http://dx.doi.org/10.1016/j.inteco.2023.01, URL https://ideas.repec.org/a/eee/inteco/v173y2023icp249-259.html.

Cerqua, A., Di Stefano, R., Letta, M., Miccoli, S., 2021a. Local mortality estimates during the COVID-19 pandemic in Italy. J. Popul. Econ. 1–29.

Cerqua, A., Di Stefano, R., Letta, M., Miccoli, S., 2021b. Local mortality estimates during the COVID-19 pandemic in Italy. J. Popul. Econ. 34 (4), 1189–1217.

Cerqua, A., Letta, M., 2022. Local inequalities of the COVID-19 crisis. Reg. Sci. Urban Econ. 92, 103752.

Checchi, D., 2000. University education in Italy. Int. J. Manpow. 21 (3–4), 177–205.

Di Pietro, G., 2004. The determinants of university dropout in Italy: A bivariate probability model with sample selection. Appl. Econ. Lett. 11 (3), 187–191.

Di Pietro, G., Cutillo, A., 2008. Degree flexibility and university drop-out: The Italian experience. Econ. Educ. Rev. 27 (5), 546–555.

Eather, N., Mavilidi, M.F., Sharp, H., Parkes, R., 2022. Programmes targeting student retention/success and satisfaction/experience in higher education: A systematic review. J. Higher Educ. Policy Manag. 1–39.

Einav, L., Levin, J., 2014. The data revolution and economic analysis. Innov. Policy Econ. 14 (1), 1–24.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874.

Friedman, J.H., 2000. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29, 1189–1232.

Friedman, J., Hastie, T., Tibshirani, R., et al., 2001. The Elements of Statistical Learning, Vol. 1. Springer series in statistics, New York.

Ghignoni, E., 2017. Family background and university dropouts during the crisis: The case of Italy. Higher Educ. 73 (1), 127–151.

Goldin, C., Katz, L., 2008. The Race Between Education and Technology. Harvard University Press, URL http://www.hup.harvard.edu/catalog.php?isbn=9780674035300.

Hughes, N., Soh, W.Y., Lawson, K., Lu, M., 2022. Improving the performance of micro-simulation models with machine learning: The case of Australian farms. Econ. Model. 115 (C), http://dx.doi.org/10.1016/j.econmod.2022.10, URL https://ideas.repec.org/a/eee/ecmode/v115y2022ics0264999322002036.html.

Jia, P., Maloney, T., 2015. Using predictive modelling to identify students at risk of poor university outcomes. Higher Educ. 70 (1), 127–149.

Johnes, G., McNabb, R., 2004. Never give up on the good times: Student attrition in the UK. Oxf. Bull. Econ. Stat. 66 (1), 23–47.

Kemper, L., Vorhoff, G., Wigger, B.U., 2020. Predicting student dropout: A machine learning approach. Eur. J. Higher Educ. 10 (1), 28–47.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S., 2018. Human decisions and machine predictions. Q. J. Econ. 133 (1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction policy problems. Amer. Econ. Rev. 105 (5), 491–495, URL https://ideas.repec.org/a/aea/aecrev/v105y2015i5p491-95.html.

Kuhn, M., 2021. CARET: Classification and Regression Training. URL https://CRAN.R-project.org/package=caret R package version 6.0-90.

Lagravinese, R., 2015. Economic crisis and rising gaps north–south: Evidence from the Italian regions. Camb. J. Reg. Econ. Soc. 8 (2), 331–342.

Lema, M.D., Vooren, M., Cannistrà, M., van Klaveren, C., Agasisti, T., Cornelisz, I., 2023. Predicting dropout in higher education across borders. Stud. Higher Educ. 1–16. http://dx.doi.org/10.1080/03075079.2023.2224818.

Li, Q., An, L., Zhang, R., 2023. Corruption drives brain drain: Cross-country evidence from machine learning. Econ. Model. 126 (C), http://dx.doi.org/10.1016/j.econmod.2023.10, URL https://ideas.repec.org/a/eee/ecmode/v126y2023ics0264999323001918.html.

Löfgren, C., Ohlsson, H., 1999. What determines when undergraduates complete their theses? Evidence from two economics departments. Econ. Educ. Rev. 18 (1), 79–88.

Modena, F., Rettore, E., Tanzi, G.M., 2020. The effect of grants on university dropout rates: Evidence from the Italian case. J. Hum. Cap. 14 (3), 343–370.

Mullainathan, S., Spiess, J., 2017. Machine learning: An applied econometric approach. J. Econ. Perspect. 31 (2), 87–106.

OECD, 2019. Education at a Glance 2015: OECD Indicators. OECD Publishing, Paris.

Oppedisano, V., 2011. The (adverse) effects of expanding higher education: Evidence from Italy. Econ. Educ. Rev. 30 (5), 997–1008.

Psacharopoulos, G., Patrinos, H.A., 2018. Returns to investment in education: a decennial review of the global literature. Educ. Econ. 26 (5), 445–458. http://dx.doi.org/10.1080/09645292.2018.1484426.

Qiu, Y., Zheng, Y., 2023. Improving box office projections through sentiment analysis: Insights from regularization-based forecast combinations. Econ. Model. 125, 106349.

Resce, G., Vaquero-Piñeiro, C., 2022. Predicting agri-food quality across space: A machine learning model for the acknowledgment of Geographical Indications. Food Policy 112, 102345.

Ripley, B., Venables, W., Ripley, M.B., 2016. Package 'nnet'. R package version 7 (3–12), 700.

Sansone, D., 2019. Beyond early warning indicators: High school dropout and machine learning. Oxf. Bull. Econ. Stat. 81 (2), 456–485.

Stinebrickner, T., Stinebrickner, R., 2012. Learning about academic ability and the college dropout decision. J. Labor Econ. 30 (4), 707–748. http://dx.doi.org/10.1086/666525, URL https://ideas.repec.org/a/ucp/jlabec/doi10.1086-666525.html.

Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Von Hippel, P.T., Hofflinger, A., 2021. The data revolution comes to higher education: Identifying students at risk of dropout in Chile. J. Higher Educ. Policy Manag. 43 (1), 2–23.

Weber, S., Péclat, M., 2017. A simple command to calculate travel distance and travel time. Stata J. 17 (4), 962–971.