FACULTY OF LAW, ECONOMICS AND FINANCE

Department of Economics and Management

Discussion Paper

2022-06

Economics

Department of Economics and Management
University of Luxembourg

Predicting dropout from higher education: Evidence from Italy

available online : https://wwwfr.uni.lu/recherche/fdef/dem/publications/discussion_papers

Marco Delogu, University of Sassari, IT
& Université du Luxembourg (Extramural Research Fellow)
Raffaele Lagravinese, University of Bari, IT
Dimitri Paolini, CRENOS & University of Sassari, IT & Core UCL, BE
Giuliano Resce, University of Molise, IT

May 2022

For editorial correspondence, please contact: dem@uni.lu
University of Luxembourg
Faculty of Law, Economics and Finance
6, Rue Richard Coudenhove-Kalergi
L-1359 Luxembourg

Predicting dropout from higher education: Evidence from Italy.*

Marco Delogu[†1], Raffaele Lagravinese[2], Dimitri Paolini[3], and Giuliano Resce[4]

[1]DISEA and CRENoS, University of Sassari and DEM, University of Luxembourg

[2]Department of Economics and Finance, University of Bari "A.Moro"

[3]DISEA and CRENoS, University of Sassari and CORE, Catholic University of Louvain

[4]Department of Economics, University of Molise

Abstract

We investigate whether machine learning (ML) methods are valuable tools for predicting students' likelihood of leaving pursuit of higher education. This paper takes advantage of administrative data covering the entire population of Italian students enrolled in bachelor's degree courses for the academic year 2013-2014. Our numerical findings suggest that ML algorithms, particularly random forest and gradient boosting machines, are potent predictors pointing to their use as early warning indicators. In addition, feature importance analysis highlights the role of the number of European Credit Transfer System (ECTS) obtained during the first year for predicting the likelihood of dropout. Accordingly, our analysis suggests that policies that aim to boost the number of ECTS gained during the early academic career may be effective in reducing drop-out rates at Italian universities.

JEL CLASSIFICATION: C53; C55; I20; KEYWORDS: Early warning system; Machine learning; Dropout; Italy

*We would like to acknowledge the participants of the seminar at University CY Cergy Paris University. The authors gratefully acknowledge financial support from: Regione Autonoma della Sardegna (Legge n. 7), DISEA, Dipartimento di Eccellenza 2018-22, and University of Sassari (fondo di Ateneo per la Ricerca 2020). We are grateful to ANVUR for sharing the data with us.

[†]Corresponding author: Marco Delogu, University of Sassari, Via Muroni 23, 07100 Italy. E-mail: mdelogu@uniss.it

1 Introduction

In recent years, the literature has extensively analyzed the determinants of university dropout in advanced countries using increasingly sophisticated empirical tools and administrative data with ever higher levels of information. The issue of university dropout, along with the NEET phenomenon,[1] represents something that should be considered very carefully. Evidently, higher drop-out rates have a detrimental impact on the overall skill composition of the workforce. In the coming years, the most desired jobs will need an increasingly advanced level of qualification that only the completion of a highly specialized university-type course can guarantee.[2] A more educated workforce would facilitate technological change and technology adoption and have a positive effect in terms of economic growth thus leading to improved efficiency (Acemoglu, 2002). Failure to complete higher education not only represents a waste of time and resources for students and their families, but it is also a misuse of public funding as long as education is usually substantially subsidized. Although there is almost unanimous consensus on the role of education in increasing one's income,[3] the percentage of students who drop out of university courses remains significantly high in many developed countries. Among the countries in the OECD area, Italy is undoubtedly an emblematic case with more worrying numbers than other developed countries. As highlighted in the OECD report (OECD, 2019), although there has been an improvement in recent years, the percentage of droppers in Italy is still among the highest in developed countries. The phenomenon of university dropout has characterized the Italian system for a long time. However, numerous reforms that have taken place over the years to increase the pool of graduates do not seem to have had the desired effect (Bratti et al., 2008; Brunori et al., 2012; Oppedisano, 2011).

This paper contributes to the literature on employing machine learning (ML) methods to develop early warning systems that predict students at risk of university dropout. Identifying groups of individuals at risk would enable universities to put in place policies to prevent students from dropping out and eventually increase the pool of graduates.

To investigate the determinants of dropout in Italy, previous work has employed standard

[1] NEET is an acronym for Not in Education Employment or Training.

[2] For instance, a recent ECB survey of leading Eurozone companies looking at 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.

[3] Psacharopoulos and Patrinos (2018) in a recent review of the literature, suggested 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).

econometric models (i.e., OLS, GLM, probit, logit, panel).[4] 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)). Nowadays, scholars are taking advantage of ML procedures to finalize public policies and predictions ((Antulov-Fantulin et al., 2021; Carrieri et al., 2021; Kleinberg et al., 2018; Mullainathan and Spiess, 2017)).

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 dropout rates. If ML show a strong ability to predict dropout 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 employed administrative data collected at a university in New Zealand to test their predictive risk model and identify students who risked dropping out. The first studies that specifically used ML algorithms primarily concerned the USA. In their work Aulck et al. (2016) analyzed the causes of dropout within the first year among students enrolled 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. Again in the USA, but analyzing a different age group, Sansone (2019) used various ML algorithms to identify the causes of dropout during the first year of high school. The author showed that schools can obtain more precise predictions by exploiting the available high-dimensional data together with ML tools such as support vector machines, boosted regression and post-LASSO. 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 found financial aid to be the main predictor of university dropout. As for Italy, at present the only work that

[4]See Aina (2013); Belloc et al. (2010); Cingano et al. (2007); Di Pietro (2004); Di Pietro and Cutillo (2008); Ghignoni (2017); Modena et al. (2020); Zotti (2015).

has used an ML approach is that of Cannistrà et al. (2021) who applied ML algorithms in the case of the Polytechnic of Milan. Their study identified previous and early academic performance as the main predictors for dropout. In particular, they found first semester results (passing or failing exams) crucial for the continuation of studies.

Our work fits into this strand of the literature and enriches it. First, 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).

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 dropout 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 144923 students for whom the relevant information was available.

Our results confirm the finding of Cannistrà et al. (2021) showing 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. To graduate Italian students need to attain a positive grade across all the set of exams that make their study plan, whereas in most other European countries, enrollment in subsequent years is conditional on students obtaining a positive average across the entire set of exams (thus meaning that some exams can be failed). Another contribution is that our paper considers a battery of algorithms. Specifically, we use four types of algorithm: (1) the least absolute shrinkage and selection operator (LASSO); (2) the random forest (RF); (3) gradient boosting machines (GBM); (4) and the neural network (NN). The ML algorithms showing the highest predictive power were the GBM and the RF; with RF performing slightly better than GBM. Finally, our findings provide additional evidence for the role of the family income, high-school grade, and high-school type. Also, RF found that 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 rest of this work is structured 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 and provides policy suggestions derived from the results obtained.

2 Institutional setting

This section highlights some peculiarities of the Italian university system in light of the objective of this study. The Eurydice network,[5] produces detailed information about the Italian University system relative to those of other European countries.[6]

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.[7] 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.[8] 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.[9]

[5]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 organised in Europe and how they work.

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

[7]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.

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

[9]These studies are organized in single cycle courses of 5-6 years, corresponding to 300-360 CF; usually they result

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.[10] 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. Interestingly, failing one attempt does not prevent further attempts from being made. Also, it is pretty 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. Second, and more interesting for readers without experience of the Italian higher education system, students who have obtained a positive grade that they are not happy with may retake the exam. It should be pointed 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

in the higher level, single-cycle, *laurea magistrale*.

[10] 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."*

can increase the grade, thereby raising his/her average, which is the critical determinant for the student's final grade. Note that in Italy final grades are crucial when candidates are competing for public administration jobs.

3 Dataset

Our data comes from ANS national registry of students enrolled in higher education institutions in Italy.[11] 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.[12]

Moreover, we excluded international students, as they are selected from a different population compared to national students and constitute a self-selected group, so that the drop-out mechanisms for them would probably be different from those that characterize domestic students. Finally, we excluded students enrolled in online universities.[13] After these exclusions our data contained information on $230,336$ students.

The next step is to differentiate between dropouts and non-dropouts. First, it should be noted that due to the peculiar characteristics of the Italian university system, contrary to Johnes and McNabb (2004), it is not possible to distinguish between voluntary and involuntarily

[11] 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).

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

[13] 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 of the students enrolled in Italian online universities are workers, therefore their determinants for dropping out of graduate studies are likely different from those students enrolled in other first-cycle programs.

university dropout.[14] 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 by the 21 March 2018, having not yet completed their degree; (3) students who changed course/university the year after their first year of enrollment; (4) students who left the Italian university system. 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. 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.

Table 1: Definition of dropout variable

Student Outcome	Number	Percentage
Enrolled but degree not yet obtained (DO_i=0)	71395	31.00
Changed course/university (DO_i=0)	41009	17.80
Degree obtained on time (DO_i=0)	88221	38.30
Left higher education (DO_i=1)	29707	12.90

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).

Our aim is 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 guided from the

[14] Involuntary drop out refers to students 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 drop out is almost impossible to define in the Italian System.

availability of data and the literature. Unfortunately, we do not have information for all features for all students. We end up with a dataset containing information on $144,904$ individuals, representing 62% of the population of students initially included.

The existing literature provides evidence that the characteristics of universities, field of study, and social and economic conditions of the students' home districts are correlated with drop-out rates, see Aina et al. (2018).

In the set of predictors we include variables capturing students' demographic information. We include the variable *Sex*, which takes value of one if the student is classified as female or zero otherwise. Descriptive statics show that unconditional to other characteristics, men leave graduate studies more compared to women.[15]

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.

In line with our assumption drop-out students earn much fewer credits during the first year than non-dropouts. Also, by computing basic descriptive statistics we find that drop-out rates are much higher for students from vocational high schools, and this finding holds for all areas (health, science, social science and humanities). Students coming from a *liceo* show a drop-out rate that is 10% lower. Conversely, students from vocational schools show a much higher drop-out rate, which reaches 21% for the science area. Accordingly, among the set of features, we include the variable HT_i, which takes a value of one if the student has earned an high school degree in a *liceo* or zero otherwise. Another important feature included is the high school grade. It is natural to expect that individuals with a low high school grade are over-represented among the dropouts, leading us to include a continuous variable capturing students' high-school grades among the drop-out determinants. Specifically, HG_i is a discrete variable and takes values in the interval $[041]$. A student enrolled in an Italian high school needs to achieve a minimum final grade of $60/100$ in order to graduate.[16]

The other two features included account for the age of the student when enrolled. Late enroll-

[15] In our dataset, 54.2% of students are female. We find that the percentage of women who leave the university (14.8%), is lower than that of men (11.2%). This finding holds for all the 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.

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

ment in Italy can be due either to grade repetition in high school or general late enrollment. Most Italian students finish high school at the age of nineteen. Accordingly, we include the variable $AGE_i = -1\,(Yearofbirth - 1995)$. However, in Italy some students can finish high school at age eighteen if they anticipated entrance at the primary school, accordingly, we include a dummy variable *Ant* which takes a value of one if the the student had an age lower than 19 or 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 enrolment. Importantly, private universities enjoy a much larger degree of freedom when setting tuition fees.[17] Accordingly, we include the variable $Tax_{i,j,c,2013}$, to reflect that the amount of tuition fees that the student had to pay during the academic year 2013-2014. Also, it is important to highlight that in Italian universities the payment of tuition fees is not upfront. Students can attend university courses without paying fees, which are normally required at the end of the academic year. Notice that if a student does not pay the tuition fees his/her exams will not be registered. As an additional proxy to family income, we include the variable $Income_o$ which is the gross average income in the student's municipality.

Other determinants of drop-out behavior include the characteristics of the field of study, the university, and some specifics of the course selected by the students. 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. Also, we include the variable $SizeCourse_{jc}$ 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

[17] In contrast to most other continental European countries, tuition fees charged by Italian public universities are not uniformly determined by the central government. According to Italian law (Decree of the President of the Republic of 25 July 1997, No. 306), the total amount of fees collected by a public university cannot exceed 20% of the funding received by the university from the MUR. Conversely, for Italian private institutions, this 20% limit does not apply, and they do charge higher fees. 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. Beine et al. (2020) reported that only private universities charged more than 2000 euro.

higher education. We include two continuous variables, $TD_{o,d}$ and $TT_{o,d}$, to capture students' off-site status.[18]

Another variable that may correlate with drop-out behavior, (see (Card, 1993)), is the the distance from the student's place of residence to the closest university. Also, we include the variable AV, which takes a value of one if the student enrolls at a university located in a different district that his/her place of residence or zero otherwise.[19] Table A2 in the Appendix, provides a brief details of definitions, data source, and remarks about all the variables employed in our analysis.

4 Methods

Every student i has an associated target binary variable DO_i (drop-out) 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) that predicts drop-out DO_i:

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

The standard routine 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, in which its predictive power is tested (Antulov-Fantulin et al., 2021; Cerqua et al., 2021). The size of these two sets must be chosen while taking in to 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 but is not generalizable (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

[18] To determine those variables, we take advantage of the STATA routine developed in Weber and Péclat (2017) and exploit information on students' home residence for all students enrolled in courses at any given university j. Then, after having obtained geographic coordinates for university j, we compute the travel distance between university j and the place of residence of any student i.

[19] Not all Italian districts host an university, although each Italian region is home to at least one.

randomly divide the database so that 80 percent for training and 20 percent 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 optimisation algorithms available through the `caret` package (Kuhn, 2021). Four different models have been analyzed:

- the Least Absolute Shrinkage and Selection Operator (LASSO): A regression statistical method that performs features selection and regularization with L1 norm to reduce over-fitting and increase prediction accuracy and interpretability (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);
- 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, 2001);
- 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).

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 are clustered in the negative class. The comparison of predicted with actual value gives rise to four possible outcomes usually reported in a table known as confusion matrix. Table 2 reports a tailored example of a confusion matrix.

Table 2: Confusion matrix

		Data	
		Dropout	Non-Dropout
Predicted	Dropout	TP	**FP**
	Non-Dropout	**FN**	TN

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

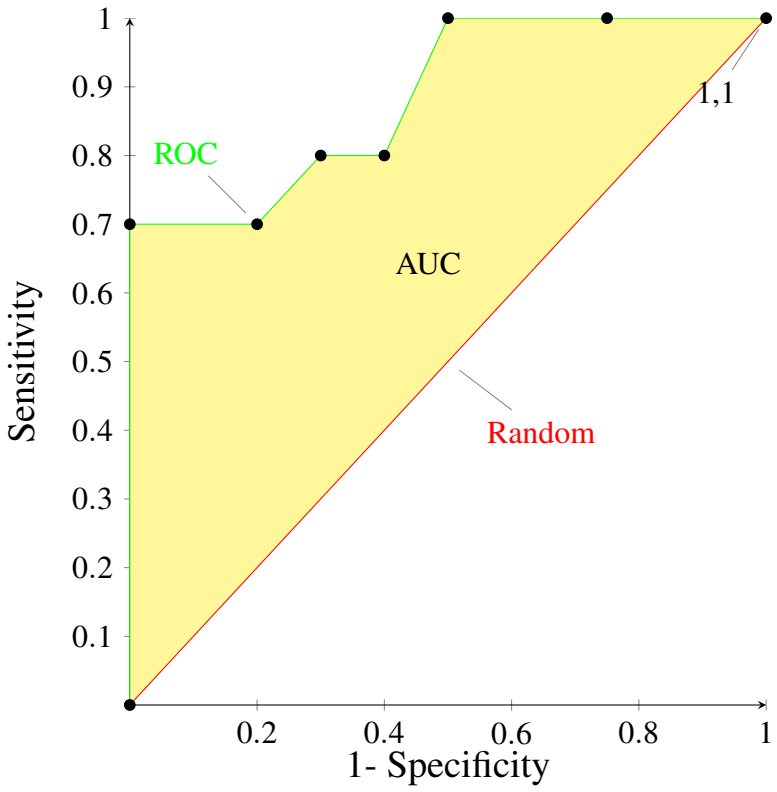
In Table 2, TP stands for true positive, and refers to students who have dropout from higher education and are correctly classified. Conversely, TN stands for true negative, referring to students who have not dropped out from higher education and are correctly classified. However, the algorithm does not always achieve the correct answer; it can make two errors, summarized at the top-right corner and bottom-left corners of Table 2. FP stands for false positive, which refers to students who are non-dropouts but classified as dropouts. FN stands for false-negative, namely students who have left higher education and are incorrectly classified as non-dropouts. Sansone (2019) provided a microfounded analysis suggesting that the ML methods, in this specific task, should aim to reduce as much as possible the number of FN errors when the policymakers' aim coincides with shrinking drop-out rates as much as possible.

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 dropout student is correctly classified, $\frac{TP}{TP+FN}$. Conversely, the specificity is the probability that successful students are classified as such, $\frac{FP}{TN+FP}$.[20] However, one main drawback is that comparing the confusion matrix across several models becomes pretty cumbersome. Consequently, in practice, the performance of classification prediction is assessed by the receiver operating characteristics (ROC) curve (Fawcett, 2006). The ROC curve shows the classi-

[20]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.

fier'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). The false-positive rate is equal to 1- specificity; it is easy to see that it is equal to the ratio between the number of students with low drop-out risk but wrongly categorized as high drop-out risk (false positive) and the total number of actual negative samples (low drop-out-risk students), $\frac{FP}{FP+TN}$. 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. Figure 1 provides an example of the ROC curve highlighting the AUC.

Figure 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.

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, (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.

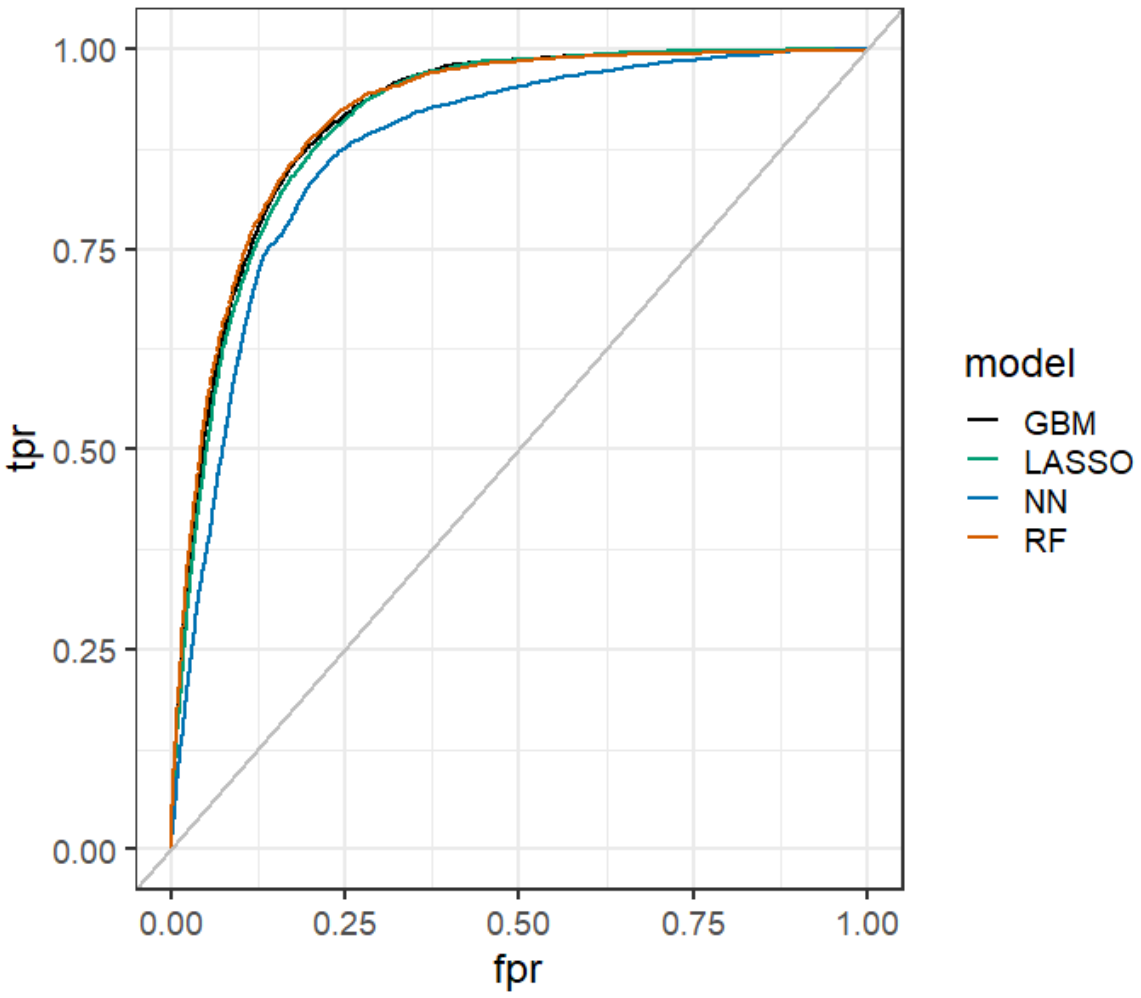
5 Results

In this section, we present the results of the model 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).

5.1 Predictability of drop-out

Figure 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 maximising 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 show lower performances.

Figure 2: ROC curve for four ML models



Models trained on 80% of observations and tested on the remaining 20%.

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 the four models used here (RF, GBM, NN, LASSO). Table 3 shows that the RF and GBM models overperform the other models in any of the metrics used: accuracy, sensitivity, specificity, detection rate, and balanced accuracy. 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 more competitive methods for structured binary tasks (Antulov-Fantulin et al., 2021; Carmona et al., 2019; Climent et al.,

2019).

Table 3: Performance of the models

	RF	GBM	NN	LASSO
Accuracy	0.898	0.895	0.876	0.891
95% CI	(0.8945, 0.9015)	(0.8912, 0.8983)	(0.8726, 0.8802)	(0.8876, 0.8948)
No information rate	0.871	0.871	0.871	0.871
P-Value [Acc $>$ NIR]	0.000	0.000	0.005	0.000
Sensitivity	0.471	0.443	0.183	0.386
Specificity	0.961	0.962	0.979	0.966
Pos pred value	0.643	0.630	0.561	0.626
Neg pred value	0.925	0.921	0.890	0.914
Prevalence	0.129	0.129	0.129	0.129
Detection rate	0.061	0.057	0.024	0.050
Detection prevalence	0.094	0.091	0.042	0.079
Balanced accuracy	0.716	0.702	0.581	0.676

Sansone (2019) argues that the most relevant metric predicting drop-out 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 at first sight such values may seem low, they are in line with or higher than the ones previously reported in the literature, (see (Kemper et al., 2020; Sansone, 2019)). As an additional empirical application, in this paper we also test the predictability of dropout within each area of study (health, science, social science, and humanities). Table 4 shows that dropout rates are predictable for three out of the four areas.[21] Restringing the attention to the algorithms best in terms of accuracy, we find that accuracy is statistically higher than the no information rate for science, social science, and humanities, while it is not significant

[21]Notice that health area mainly contains students enrolled in nursing studies. In Italy, each university (for health courses) can accept a set number of students and places are allocated according to the results of an entrance test, for which we have no information. We think that the low dropout rates are mainly due to the high employability of those students once they complete their studies.

in the case of health.

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

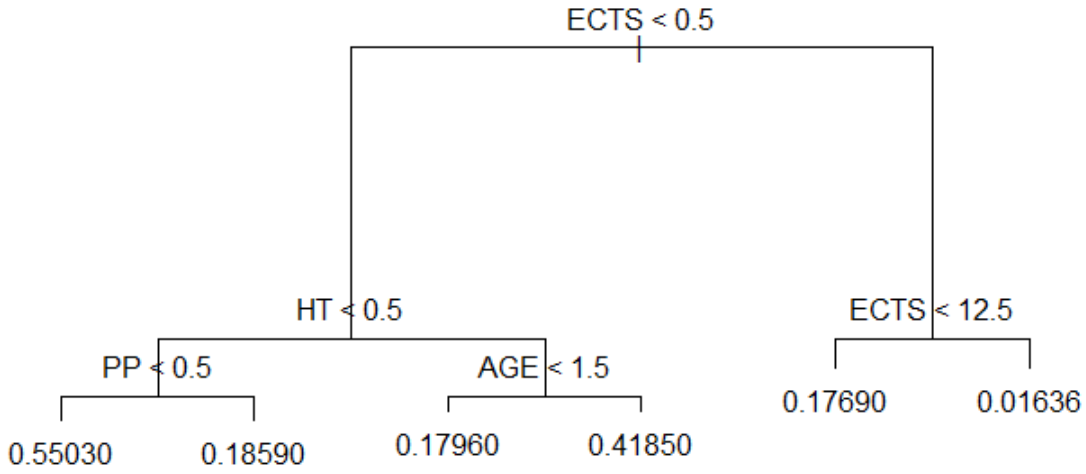
We report goodness of fit measures for the best model (among RF, GBM, NN, and LASSO) in terms of AUC for each area of study.

Notice that for both social science and humanities, our best models produce a true positive rate (sensitivity) larger than 50%. Therefore, clustering the data among the study areas increases the prediction power of regression tree methods.

5.2 Features importance

This section shows the most important features in the prediction task. The previous section showed that the two best-performing models (RF and GBM) are based on combinations of different regression trees. Although the standard regression tree has low predictive power, it highlights the most critical variables in the prediction task. We consider a standard regression tree with tree branches.

Figure 3: Regression tree over the whole sample

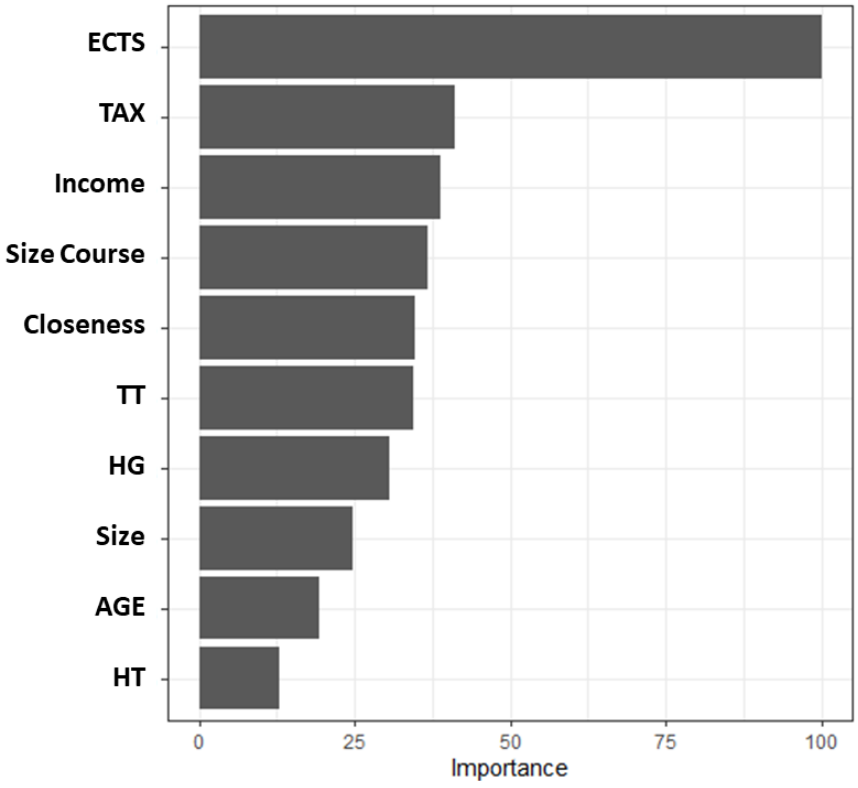


From Figure 3 it can be observed that the number of ECTS earned by the student at the end of the academic year is the most critical factor for explaining dropout, as this feature is on the top of the tree. For students who did not earn any ECTS, the type of high school (*liceo* vs. *non liceo*) and the type of university (public vs. private) 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 enrolls 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 learn that the dropout probability is almost equal to zero for students who earned more than 12.5 ECTS during their first year.

Figure 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 Figure 3 the most important feature is the number of ECTS earned by

the students at the end of the academic year. 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 place of residence and the university, physical distance between the student 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. Our feature importance analysis is in line with the results reported in the literature, (see (Aina et al., 2018)). Also, 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.

Figure 4: Feature importance for predicting dropout: The first 10 important features in RF



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

This section tests the predictability of dropout splitting the sample in northern (102.702 observations) and southern (40.012 observations) regions. 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.[22] 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). Notice that sensitivity is much larger for southern students when they are considered separately, with ML methods better able to detect those at risk of dropout in southern universities, the ones where dropout is more common. For the same ML algorithms, we conduct feature importance analysis (Table 5).

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

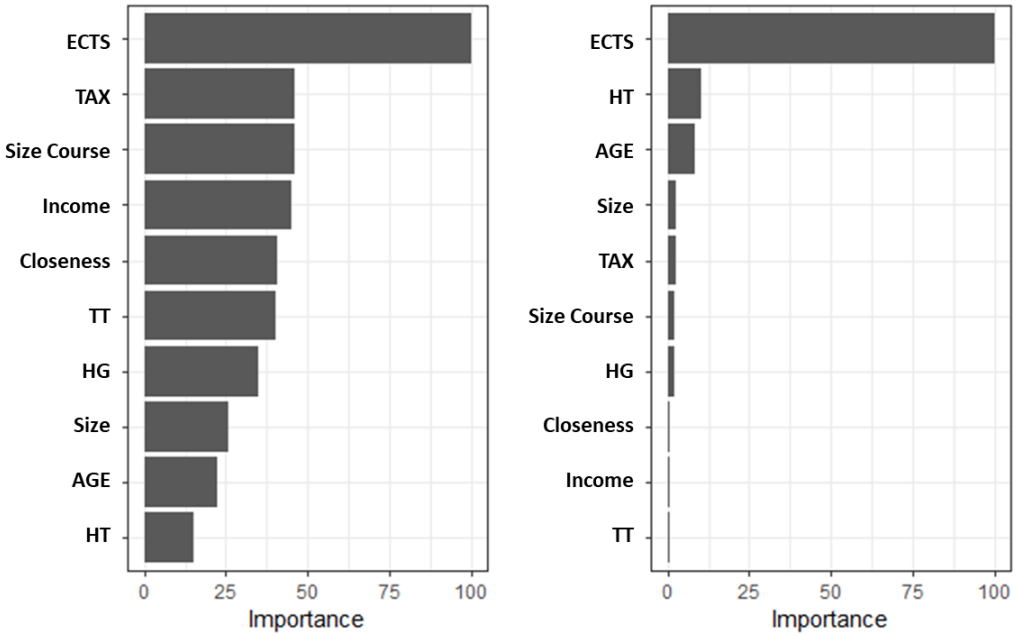
	North (RF)	South (GBM)
Accuracy	0.9080	0.8688
95% CI	(0.9039, 0.9119)	(0.8612, 0.8761)
No information rate	0.8888	0.8258
P-Value [Acc $>$ NIR]	0.0000	0.0000
Sensitivity	0.4210	0.5273
Specificity	0.9689	0.9408
Pos pred value	0.6292	0.6528
Neg pred value	0.9304	0.9042
Prevalence	0.1113	0.1742
Detection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

We report goodness of fit measures for the best model (among RF, GBM, NN, and LASSO) in terms of AUC.

[22] Table 5 shows the models with the best accuracy for both subsamples. Additional results are available upon request.

From Table 5 certain degree of heterogeneity emerges in terms of feature importance. Figure 5 shows that while ECTS is the most important feature in both the 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. Our results suggest that, in line with the results of Stinebrickner and Stinebrickner (2012), 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 from higher education. Notice that the low predictive power of the *TAX* variable when we consider southern students separately can be partly explained by the fact that several grants are available to students from poor households residing in southern regions.

Figure 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)



6 Conclusions

This paper investigates whether ML methods are suitable for identifying higher education students at risk of dropping out. 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. Although model accuracy slightly increased with decision trees, we reported a substantial increase in terms of sensitivity, meaning that decision trees algorithms are much better for correctly identifying students at high risk of dropout.
We also conducted a features importance analysis. In line with the literature, we showed that the strongest predictor of drop-out behavior is the number of 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. The feature analysis also confirmed the importance of family background on dropout behavior, with taxes (function of family income in Italy), high school type, and high school grade among the most important predictors. In addition, we performed the prediction task considering separately northern and southern universities. Like the previous analysis, decision trees algorithms were the ones giving the best results in terms of accuracy and sensitivity. However, differences arose when conducting the feature analysis, with family income resulting in greater importance when predicting the drop-out decision for northern students.

Our analysis has policy implications. First, dropout is predictable, and ML algorithms (specifically regression tree methods) can be used to identify students at risk. By knowing this subset of students, universities might put in place policies, such as remedial courses, aimed at preventing university dropouts. 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 (see Section 2) 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.

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 programmes for university orientation, retention/success, and satisfactory experience should be carefully investigated (Eather et al., 2022).

References

Acemoglu, D. (2002). Directed Technical Change. *The Review of Economic Studies*, 69(4):781–809.

Aina, C. (2013). Parental Background and University Dropout in Italy. *Higher Education*, 65(4):437–456.

Aina, C., Baici, E., Casalone, G., and Pastore, F. (2018). The Economics of University Dropouts and Delayed Graduation: A Survey. GLO Discussion Paper Series 189, Global Labor Organization (GLO).

Antulov-Fantulin, N., Lagravinese, R., and Resce, G. (2021). Predicting Bankruptcy of Local Government: A Machine Learning Approach. *Journal of Economic Behavior & Organization*, 183:681–699.

Atzeni, G., Deidda, L., Delogu, M., and Paolini, D. (2022). *Drop-Out Decisions in a Cohort of*

Italian Universities, volume in Teaching, Research, and Academic Careers edited by Checchi D., Jappelli T., and Uricchio F. Springer.

Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting Student Dropout in Higher Education. *arXiv preprint arXiv:1606.06364*.

Beine, M., Delogu, M., and Ragot, L. (2020). The Role of Fees in Foreign Education: Evidence from Italy. *Journal of Economic Geography*, 20(2):571–600.

Belloc, F., Maruotti, A., and Petrella, L. (2010). University Drop-Out: An Italian Experience. *Higher Education*, 60(2):127–138.

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

Bratti, M., Checchi, D., and 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. *Machine Learning*, 45(1):5–32.

Brunori, P., Peragine, V., and Serlenga, L. (2012). Fairness in Education: the Italian University Before and After the Reform. *Economics of Education Review*, 31(5):764–777.

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2021). Early-Predicting Dropout of University Students: An Application of Innovative Multilevel Machine Learning and Statistical Techniques. *Studies in Higher Education*, pages 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.

Card, D. (2001). Estimating the Return to Schooling: Progress on some Persistent Econometric Problems. *Econometrica*, 69(5):1127–1160.

Carmona, P., Climent, F., and Momparler, A. (2019). Predicting Failure in the US Banking Sector: An Extreme Gradient Boosting Approach. *International Review of Economics & Finance*, 61:304–323.

Carrieri, V., Lagravinese, R., and Resce, G. (2021). Predicting Vaccine Hesitancy from Area-Level Indicators: A Machine Learning Approach. *Health Economics*, 30(12):3248–3256.

Cerqua, A., Di Stefano, R., Letta, M., and Miccoli, S. (2021). Local Mortality Estimates during the COVID-19 Pandemic in Italy. *Journal of Population Economics*, pages 1–29.

Checchi, D. (2000). University Education in Italy. *International Journal of Manpower*, 21(3-4):177–205.

Cingano, F., Cipollone, P., et al. (2007). *University Drop-Out: The Case of Italy*, volume 626. Banca d'Italia Roma.

Climent, F., Momparler, A., and Carmona, P. (2019). Anticipating Bank Distress in the Eurozone: An Extreme Gradient Boosting Approach. *Journal of Business Research*, 101:885–896.

Di Pietro, G. (2004). The Determinants of University Dropout in Italy: A Bivariate Probability Model with Sample Selection. *Applied Economics Letters*, 11(3):187–191.

Di Pietro, G. and Cutillo, A. (2008). Degree Flexibility and University Drop-Out: The Italian Experience. *Economics of Education Review*, 27(5):546–555.

Eather, N., Mavilidi, M. F., Sharp, H., and Parkes, R. (2022). Programmes Targeting Student Retention/Success and Satisfaction/Experience in Higher Education: A Systematic Review. *Journal of Higher Education Policy and Management*, pages 1–39.

Einav, L. and Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*, 14(1):1–24.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861–874.

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

Friedman, J. H. (2001). Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*, pages 1189–1232.

Ghignoni, E. (2017). Family Background and University Dropouts during the Crisis: The Case of Italy. *Higher Education*, 73(1):127–151.

Jia, P. and Maloney, T. (2015). Using Predictive Modelling to Identify Students at Risk of Poor University Outcomes. *Higher Education*, 70(1):127–149.

Johnes, G. and McNabb, R. (2004). Never Give up on the Good Times: Student Attrition in the UK. *Oxford Bulletin of Economics and Statistics*, 66(1):23–47.

Kemper, L., Vorhoff, G., and Wigger, B. U. (2020). Predicting Student Dropout: A Machine Learning Approach. *European Journal of Higher Education*, 10(1):28–47.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction Policy Problems. *American Economic Review*, 105(5):491–495.

Kuhn, M. (2021). *CARET: Classification and Regression Training*. R package version 6.0-90.

Löfgren, C. and Ohlsson, H. (1999). What Determines When Undergraduates Complete their Theses? Evidence from Two Economics Departments. *Economics of Education Review*, 18(1):79–88.

Modena, F., Rettore, E., and Tanzi, G. M. (2020). The Effect of Grants on University Dropout Rates: Evidence from the Italian Case. *Journal of Human Capital*, 14(3):343–370.

Mullainathan, S. and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 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. *Economics of Education Review*, 30(5):997–1008.

Psacharopoulos, G. and Patrinos, H. A. (2018). Returns to Investment in Education: a Decennial Review of the Global Literature. *Education Economics*, 26(5):445–458.

Ripley, B., Venables, W., and 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. *Oxford Bulletin of Economics and Statistics*, 81(2):456–485.

Stinebrickner, T. and Stinebrickner, R. (2012). Learning about Academic Ability and the College Dropout Decision. *Journal of Labor Economics*, 30(4):707–748.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Von Hippel, P. T. and Hofflinger, A. (2021). The Data Revolution Comes to Higher Education: Identifying Students at Risk of Dropout in Chile. *Journal of Higher Education Policy and Management*, 43(1):2–23.

Weber, S. and Péclat, M. (2017). A Simple Command to Calculate Travel Distance and Travel Time. *The Stata Journal*, 17(4):962–971.

Zotti, R. (2015). Should I Stay or Should I Go? Dropping Out from University: An Empirical Analysis of Students' Performances. In *Youth and the Crisis*, pages 71–88. Routledge.

A Appendix

A.1 Descriptive Statistics

The table below reports the region, name, legal status, and drop-out rate (as defined in Section 3) for each university in the dataset.

Region	University	Legal status	Dropouts (%)
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 government dependent	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 government-dependent	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 government dependent	10.9
UMBRIA	University for Foreigners Perugia	public	14.0
UMBRIA	University of Perugia	public	15.3
VALLE D'AOSTA	Università della Valle d'Aosta	private government dependent	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

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

Table A2: Data sources and definitions

Variable	Definition	Source	Remarks
Dropout $\left(D_{\{i,\}}\right)$	Dummy variable that takes value of one when the student drops out from the course/university or zero otherwise.	ANS data; our computation.	
$Area_{i,a}$	The subscript a captures the area of study (health, science, social science, humanities). Accordingly we build four dummy variables.	ANS data.	
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*; $AGE_i = -1\,(Yearofbirth_i - 1995)$.	ANS data; our computation.	Note that in Italy, students usually finish high school at the age of 19.
HT_i	Dummy variable that captures the type of high school attended by student 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.

Continued on next page

Table A2 – continued from previous page

Variable	Definition	Source	Remarks
G_i	Dummy variable that captures the gender of the student *i*. Takes a value of one for male or zero otherwise.	ANS data.	
$OD_{i,u,o}$	Dummy variable that takes a value of one when the students enrolls in a university not located in his/her district of residence.	ANS data.	
$TD_{i,u,o}$	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.	Our computation employed the routine developed by Weber and Péclat (2017), one unit is equal to 100km.
$TT_{i,u,o}$	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.	Our computation employed the routine developed by Weber and Péclat (2017), which determines the shortest path between locations, accounting for the means to transport.
AV	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.
Closeness	Continuous variable that measures the distance from the student's place of origin and the nearest university.	ANS data and ETER data, our computation.	We computed the distances employing the routine developed in Weber and Péclat (2017) and then determining for each place of residence the closest university.

Continued on next page

Table A2 – continued from previous page

Variable	Definition	Source	Remarks
$TAX_{i,j,c,2013}$	Amount of taxes that university j charged to student i enrolled at course c during the academic year 2013-2014.	ANS data.	Note that in Italian public universities, tuition fees should not be paid upfront.
$Income_{o,i}$	Average gross income in the place of origin, o of student i.	Italian Ministry of Economics and Finance.[23]	Row data are taken from the fiscal declaration data set available at municipal level. Original information is split into eight classes of gross income; we use class figures to estimate the average income, as in recent papers using the same dataset by Antulov-Fantulin et al. (2021) and Carrieri et al. (2021).
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.	
$ECTS_{i,2013}$	Number of ECTS earned by student i at the end of the academic year 2013-2014.	ANS data.	

[23] https://www1.finanze.gov.it/finanze/paginadichiarazioni/public/dichiarazioni.php