Social Indicators Research (2021) 156:341–362
https://doi.org/10.1007/s11205-019-02249-y

ORIGINAL RESEARCH

# A Statistical Analysis of Factors Affecting Higher Education Dropouts

**Paola Perchinunno[1] · Massimo Bilancia[2] · Domenico Vitale[3]**

Accepted: 9 December 2019 / Published online: 19 December 2019


## Abstract

One of the most significant indicators for assessing the quality of university careers is the dropout rate between the first and second year. Both literature on the subjects and the results that emerged from numerous specific investigations into the dropouts of the university system, showed the crucial importance of this junction between the first and the second year. Reasons for dropping out can be quite varied, ranging from incorrect and/or insufficient prospective student orientation, the willingness or need to find a job as quickly as possible, to a lack of awareness of not being able to cope with a particular course of study rather than another. In this paper we focus specifically on the problem of dropouts in Italy, addressing it from a dual point of view. At an aggregate level, the analysis deals with dropout rates in Italy between the first and second year, in order to identify the main trends and dynamics at the national level. Subsequently, we analyze individual-level data from the University of Bari Aldo Moro, aiming to identify the most important contributing factors. This individual-level approach has emerged over recent years, and is generally known as 'Educational Data Mining', focused on the development of ad hoc methods that can be used to discover regularities and new information within databases from contexts related to education. Using supervised classification methods, we are able to identify retrospectively the profile of students who are most likely to dropout.

**Keywords** Dropout rates · University careers · Data science · Machine learning

✉ Massimo Bilancia
massimo.bilancia@uniba.it; massi.bilancia@gmail.com

Paola Perchinunno
paola.perchinunno@uniba.it

Domenico Vitale
domvit@unitus.it

[1] Department of Business and Law Studies (DEMDI), University of Bari Aldo Moro, Largo Abbazia di Santa Scolastica n.53, 70124 Bari, Italy

[2] Ionian Department of Law, Economics and Environment, University of Bari Aldo Moro, Via Lago Maggiore angolo Via Ancona, 74121 Taranto, Italy

[3] Department for Innovation in Biological, Agro-food and Forest Systems, University of Tuscia, Via San Camillo de Lellis, 01100 Viterbo, Italy

## 1 Introduction

To assess the quality of university careers, one of the most significant indicators is the dropout rate between the first and second year, defined as the percentage change between the number of students enrolled in the second year and that of freshman new students in the previous year. Both literature on the subjects and the results that emerged from numerous specific investigations into the dropouts of the university system showed the crucial importance of this junction between the first and the second year, during which the great majority of dropouts or decisions to transfer to another course of study occurs (Tinto 1975; Johnson 1997; Paura and Arhipova 2014). Reasons for dropping out can be quite varied, ranging from incorrect and/or insufficient prospective student orientation, the willingness or need to find a job as quickly as possible, to a lack of awareness of not being able to cope with a particular course of study rather than another. Dropping out is not necessarily a definitive condition; those who have abandoned can decide to change their mind and resume their studies after a certain period of time, at the same university or even at another university.

According to Eurostat data, in 2016 more than 3 million young Europeans dropped out of university. In the ranking of the EU countries with the highest number of dropouts, Eurostat ranks France 1st (with a third of the total number of dropouts), followed by Italy with a total dropout rate of 15.8%, the third place being the United Kingdom, with 12%. According to Eurostat, 24% of students, aged between 20 and 35, dropout of university motivated by desire to enter the labour market. Much research has tried to explain the determinants of these data. For example, Smith and Naylor (2001) studied the risk of dropout in a cohort of UK university students, concluding that the likely causes were: the extent of prior academic preparedness and social integration at university, as well as the unemployment rate in the county of prior residence. For Murray (2014), financial aid and residence-based accommodation were also found to help students who would eventually graduate, while Araque et al. (2009) point out that students with weak educational strategies and without persistence to achieve their aims in life have low academic performance and a high risk of dropping out. In general, the educational background is advocated as a main influence, along with some individual characteristics of the student (Montmarquette et al. 2001).

Literature examining data related to the Italian higher education systems is extensive as well. For example, the empirical analysis conducted by Belloc et al. (2010) unveils that lower income class (ISEE< 10,000 €) drop-out less likely than rich ones, probably due to financial pressures, and that the higher the number of years between the secondary education diploma and the enrollment in the university the lower the dropping-out probability, as adult student (often workers) have stronger motivations to conclude the degree course. Surprisingly, they found that the higher the secondary school final mark, the higher the probability of university withdrawal. The authors interpreted this result as a consequence of the fact individuals with a high educational background are more sensitive to a low performance at the university, even though this result has not been confirmed by most papers on subject. For example, Di Pietro and Cutillo (2008) found that the high school diploma score has been shown as an important predictor of retention, in the sense that students with an higher diploma score are less likely to drop out. However, we found an effect similar to that reported by Belloc et al. (2010). We will propose a simple explanation of why this result should not be considered in contraddiction to the inverse relationship, as estimated on aggregated data, exisisting between secondary school final grade and abandonment rates.

Another interesting study is Cipollone and Cingano (2007), which show that the dropout probability is decreasing in father's years of formal education. Other studies carried out in relation to the Italian experience confirm the presence of a mix of endogenous/exogenous factors that, directly or inversely, are strongly correlated to the risk of dropout. Among those of particular relevance are: the chosen Study Program has a limited number of students, the quality of the freshman orientation programs, the number of students attending the courses and the perceived self-efficacy in the organization of individual study (Di Pietro 2004; Belloc et al. 2011; Burgalassi et al. 2016; Meggiolaro et al. 2017) . This body of literature show how dropouts of students are not due to a single factor that can be taken in isolation.

In the light of this complex picture, we want to contribute further on the problem of dropouts in Italy, addressing it from a dual point of view. At the aggregate level, the analysis deals with dropout rates in Italy between the first and second year, in order to identify the main trends and dynamics at the national level. Subsequently, we analyze individual-level data from the University of Bari Aldo Moro, aiming to identify the most important contributing factors. While the first approach has its own importance to facilitate the identification of the most appropriate policy guidelines to reduce dropout rates in future cohorts, the latter has emerged over recent years, and is generally known as 'Educational Data Mining' (EDM). The two approaches are closely linked, firstly because it is important to verify to what extent the dynamics valid at national level, based on aggregated data, are confirmed when we consider individual data of students enrolled on specific Universities or degrees. Secondly, individual data analysis aims at predicting the probability of dropout for each student, and is largely inspired by the churn analysis used in many marketing studies. The churn or attrition rate, is any estimate of the number of individuals who leave a certain group at a defined time interval. The churn analysis techniques aim to identify these individuals early, in order to implement actions at an individual level that increase the retention rate, thus countering dropouts (Ismail et al. 2015; Khodabandehlou and Zivari Rahman 2017). Therefore, we have two apparently distinct levels of analysis, but which actually share a common goal.

The Data Mining process, also known as 'Knowledge Discovery in Databases' (KDD), consists of the automatic discovery through appropriate algorithms of new and potentially useful information hidden within large amounts of data. The EDM is precisely focused on the development of ad hoc methods that can be used to discover regularities and new information within databases from contexts related to education, aimed at better understanding the individual students and the environments within which this instruction is provided, as well as their relation to the expected performance and objectives (Baker and Yacef 2009; Miguéis et al. 2018). The analysis and use of supervised classification algorithms that predict the performance of future students based on historical data is part of that discipline generally known as 'Machine Learning' (ML; Mitchell 1997; Ghahramani 2015). The nature of these methods, and their relationship to classical statistical inference, is discussed in more depth in Sect. 4.

The paper is organized as follows. In Sect. 2 we analyze national data from the National Agency for the Evaluation of the University and Research System (ANVUR) and National Student Registry (ANS). In particular, we analyze aggregate trends and patterns of university dropout rates between the first and second year. In Sect. 3, this aggregated assessment is narrowed to the data from the University of Bari Aldo Moro, in order to facilitate comparisons with the national dynamics. Section 4 concerns with individual profiles of students who dropout. We first analyze in more depth the concept of EDM, distinguishing between purely predictive and retrospective analyses. Then, using two classification algorithms, we seek to identify the most important variables in explaining dropouts. This process is conducted either in-sample or out-of-sample, on a predictive basis: the interplay of these two point of views provides useful informations to identify the students who are most likely to dropout. Section 5 contains a brief discussion of the results and suggests the way forward for future research.

## 2 The Dropout Rate in Italy

At the aggregate level, the National Agency for the Evaluation of the University and Research System (ANVUR) monitors the performance of the university system using the data of the National Student Registry (ANS). On the basis of these data, it is also possible to monitor, year by year, the number of dropouts at any resolution level. In particular, our analysis is focused on the following two indices:

1. University dropout rate between the first and second year of the course, concerning students who, in the transition to the second year, leave the system, being no longer enrolled in any course.
2. Mobility between the first and second year of the course: it occurs when the continuation of studies takes place in another course of study, either of the same or of another university (transfer).

From the data of the last ANVUR 2018 Report on the State of the University System, it emerges that in the bachelor's degrees the percentage of dropouts between the first and second year in the 2015/2016 cohort is 12.2%. Significantly lower dropout rates are recorded in the single cycle master's degrees (combined bachelor + master), at 7.5% in the 2015/2016 cohort, and in master's degrees, which reach 5.9%. As clearly shown in Fig. 1, the dropout rates are decidedly lower than in the previous cohorts, showing a reduction of 4% points for the bachelor's degrees from 2006/07 to 2015/2016 and about 2% points for the others. However, although the data on the most recent student cohorts show a slight

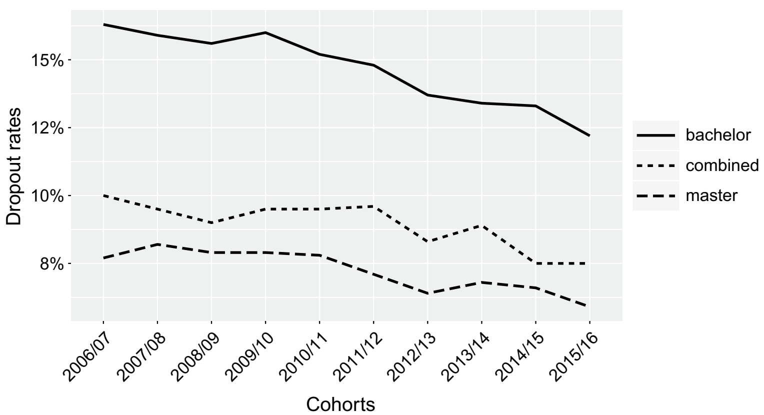


**Fig. 1** Time series of Italian university dropout rates, disaggregated by type of degree (cohorts from 2006/07 to 2015/2016). The label 'combined' refers to combined bachelor + master's degrees (single-cycle master's degrees). *Source*: National Student Registry of MIUR-Cineca

improvement, the phenomenon of dropouts must still be considered significant (Carletti 2018). The strengthening of the government policies implemented so far to combat early dropouts appears to be an inescapable necessity to comply with the Europe 2020 strategy, which sets the target dropout rate at no more than 10%.

The downward trend in dropout rates that we have just highlighted also characterizes the data broken down by scientific area. Even when we disaggregate the courses of study by CUN scientific area (CUN = Consiglio Universitario Nazionale, Italian National University Council), a general improvement emerges in recent years; in those few cases where there is an increase in the dropout rate, this increase is of little relevance and refers to a low initial level. For the last cohort of enrolled students (2015/2016), the dropout rate is relatively high in Area 01 (Mathematics and Informatics; 16.8%), Area 04 (Earth Sciences 16.4%), Area 7 (Agricultural and Veterinary Sciences; 17.1%) and Area 12 (Legal Sciences % 19.8) (see Fig. 2). The percentage of dropouts in Area 12 is particularly noteworthy and alarming, and goes together with the significant reduction in the number of students enrolled in law degrees that has occurred in recent years (according to ANVUR data, −38% from 2006 to 2018). This decrease continues to persist, as the percentage over the total number of enrolled students of the 2017/2018 cohort reduced from 9.3 to 7.2% (Carci 2018).

Further differences are found at the geographical level, as at least three points of difference are observed between the North and the South of the country; in fact, bachelor's degrees have a dropout rate of 14.3% in the south compared to 10.7% in the North. The same pattern is present in the case of single-cycle master's degrees, with a dropout rate ranging from 9.5% for Southern universities to 6.0% for Northern universities, as well as in the case of master's degrees (from 7.2 to 4.9%; see Fig. 3). This data is a further confirmation that there is no real convergence in objectives and performances between the universities of the North and those of the South of Italy. An interesting and updated analysis of Italy's educational North-South divide is contained in the OECD Skills Strategy Italy 2017 report (OECD 2017).

Considering those who continue the course of study, it is interesting to understand whether the students between the first and second year continue to attend the same course or transfer to another course in the same university, or even transfer to another university.

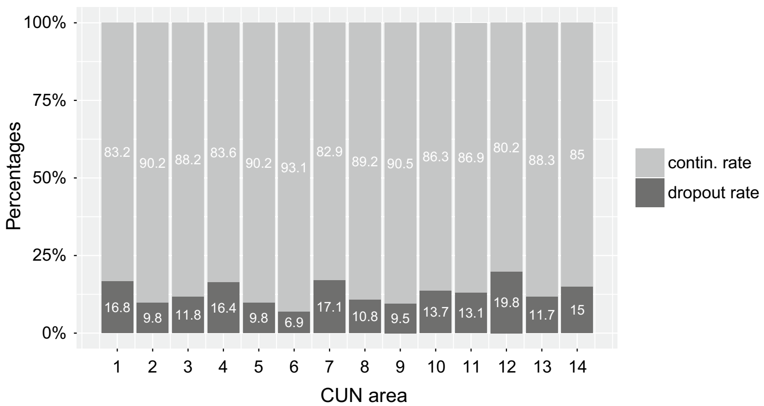


**Fig. 2** Italian university dropout rates, disaggregated by CUN scientific area (cohort 2015/2016). *Source*: National Student Registry of MIUR-Cineca

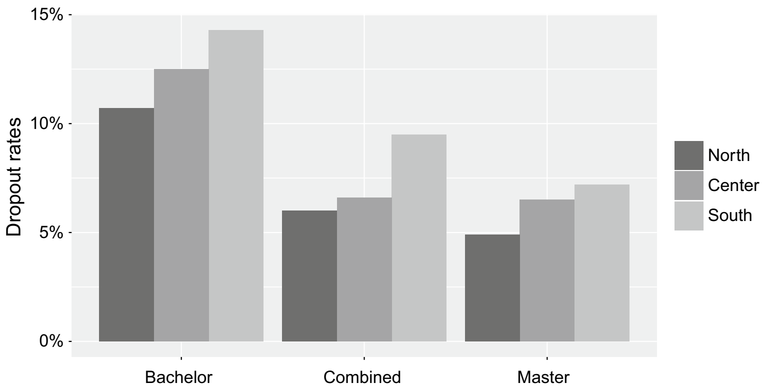


**Fig. 3** Italian university dropout rate between the first and second year of the course of study, disaggregated by geographical area (cohort 2015/2016). The label 'combined' refers to combined bachelor + master's degrees (single-cycle master's degrees). *Source*: National Student Registry of MIUR-Cineca

In the bachelor's and single-cycle master's degrees the continuation involves 73.2% and 77.4% respectively, while a course of study transfer between 1st and 2nd year involves approximately 15% of the registered students (Table 1). Among those who transfer while attending a bachelor's or a single cycle master's degree, about half transfer to another university; in the bachelor's degree, students who do not move to another university prevail slightly (7.7%), while transfers to another university are more frequent (8.2%) for single-cycle master's degrees. In the master's degree courses we observe negligible percentages.

The analysis of the data disaggregated by CUN scientific area is quite interesting. In the case of bachelor's degrees and for cohort of matriculations analyzed (2015/2016, see Fig. 4), the percentages of those who continue in another course in the same university are high, in particular in Areas 3 and 4 (Chemistry and Earth Sciences; about 20%) and Area 5 (Biological Sciences, 14.2%). Those who instead move to different course in a different university are more present in Area 3 (Chemistry; 12.1%) and Area 5 (Biological Sciences; 15.7%). With particular reference to the latter Area, transfers are likely to be related to

**Table 1** Outcome in the transition between the 1st and the 2nd year of the course, by type of course and type of continuation (cohort 2015/2016). *Source*: National Student Registry of MIUR-Cineca

| Enrolled | Bachelor<br>239,727 | Combined<br>34,908 | Master<br>108,647 |
|---|---|---|---|
| *Results between 1st and 2nd year* (%) | | | |
| Dropout | 12.2 | 7.4 | 6.2 |
| Continuations | 87.8 | 92.6 | 93.8 |
| *Continuations* (%) | | | |
| Same course of study | 73.2 | 77.4 | 91.9 |
| Different course of study | 14.6 | 15.2 | 1.9 |
| *Continuations in a different course of study* (%) | | | |
| Same university | 7.7 | 7.0 | 0.8 |
| Different university | 6.9 | 8.2 | 1.1 |

The label 'combined' refers to combined bachelor + master's degrees (single-cycle master's degrees)

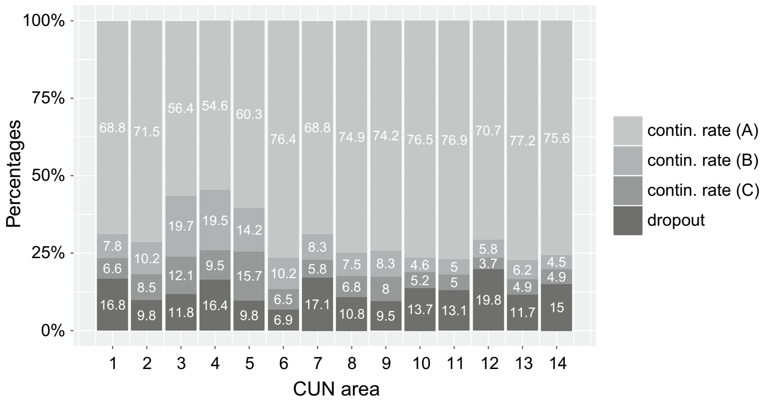


**Fig. 4** University dropout and continuation rate in the transition between the 1st and the 2nd year of the course for bachelor's degrees, by CUN area (2015/2016 cohort). Contin. rate (A) = continuation rate in the same course of study in the same university. Contin. rate (B) = continuation rate in a different course of study in the same university. Contin. rate (C) = continuation rate in a different course of study in a different university. *Source*: National Student Registry of MIUR-Cineca

those students who have failed the entrance exam for medical courses for a limited number of student (numerus clausus), and have enrolled in courses of study of this CUN Area with the intention of transferring on enrollment in the second year. In both master's and single-cycle master's degrees the variations among CUN areas are much more limited, and no significant differences emerge.

## 3 Statistical Analysis of the Dropout Rate of the Students of the University of Bari

### 3.1 Data Structure

The University of Bari Aldo Moro is one of the largest Italian universities, based in Apulia, Southern Italy. In the last academic years (2018/2019) the university population amounted to around 45,000 units; the students enrolled in the 2015/2016 academic year were approximately 11,000 (by summing up students enrolled in bachelor's, single-cycle master's and master's degree courses). Data on the university student population were collected from the National Registry Students (ANS) in March 2019. The two indicators considered in this section are the following:

1. University dropout rate between the first and second year of the course: it is calculated including those students who, in the transition to the second year, leave the University of Bari, being no longer enrolled in any course offered by the University of Bari.
2. Mobility between the first and second year of the course (course transfers): it is calculated on the number of students who transfer to another course offered by the University of Bari, either in the same degree class or in another class.

It is necessary to highlight that the dropout rate considered in this paragraph refers to all those who do not enroll in the second year at the University of Bari. Thus, it includes not only those who leave their university studies, but also those who move to another university. Therefore, it must be compared cautiously with the national dropout rate calculated in the previous paragraph, which does not include the total number of transfers to another university.

## 3.2 Dropout Rates Between First and Second Year

Our analysis refers only to dropouts between the first and second year of bachelor's and single-cycle master's degrees, as the dropouts of master's degrees are negligible compared to the total number of students. Moreover, as we already pointed out above, the dropout rate between the first and the second year considered here includes those who leave their university studies as well as those who move to another university. Therefore, in this section transfers to another university will be considered as true dropouts.

As far as the dropout rate is concerned, there are apparent differences between the two degrees (Table 2). In fact, we have a difference of about 5% points between bachelor's and single-cycle master's degrees (21.8% vs 16.4%). We also have substantial percentages among those who change course of study, choosing to leave for a course in another degree class (9.7% for bachelor's vs 11.6% for single-cycle master's degrees).

Analyzing the university dropout rates of the bachelor's degrees by the scientific area of the course of study (CUN area), it emerges as for the cohort of matriculations analyzed (academic year 2015/2016) the percentage of dropouts is relatively high in Area 04 (Earth Sciences; 39,5%) and Area 12 (Law Studies; 36.5%), see Table 3. In the case of single-cycle master's degree (Table 4), the highest dropout rates are found in Area 07 (Agricultural and Veterinary Sciences; 20%) and in Area 12 (Law Studies; 20.70%). Despite limitations in comparability that we have highlighted, these results are entirely in line with what we have obtained at the national level.

We have also analyzed the variation in abandonment rates as a function of some explanatory variables, such as:

– Gender;

**Table 2** Outcome in the transition between the 1st and the 2nd year of the course, by type of course and type of continuation (cohort 2015/2016, University of Bari Aldo Moro). *Source*: National Student Registry of MIUR-Cineca

| | Bachelor | | Combined | | Total | |
|---|---|---|---|---|---|---|
| | abs. | % | abs. | % | abs. | % |
| Continuations (A) | 3924 | 65.81 | 964 | 69.80 | 4888 | 66.56 |
| Continuations (B) | 158 | 2.65 | 31 | 2.24 | 189 | 2.57 |
| Continuations (C) | 579 | 9.71 | 160 | 11.59 | 739 | 10.06 |
| Dropouts and transfers to another university | 1302 | 21.83 | 226 | 16.36 | 1528 | 20.81 |
| Enrolled | 5963 | 81.20 | 1.381 | 18.80 | 7344 | 100.00 |

The label 'combined' refers to combined bachelor+master's degrees (single-cycle master's degrees). Continuations (A) = continuations in the same course of study of the University of Bari Aldo Moro. Continuations (B) = continuations in a different course of study of the same degrees class at the University of Bari Aldo Moro. Continuations (C) = continuations in a different course of study of another degree class at the University of Bari Aldo Moro

**Table 3** Outcome in the transition between the 1st and the 2nd year of the course for bachelor's degree by CUN scientific area (cohort 2015/2016, University of Bari Aldo Moro). *Source*: National Student Registry of MIUR-Cineca

| CUN area | Dropouts (%) | Contin. (B) (%) | Contin. (C) (%) |
|---|---|---|---|
| 01 - Math. and informatics | 22.80 | 2.83 | 5.10 |
| 02 - Physics | 27.27 | 3.03 | 15.15 |
| 03 - Chemistry | 17.68 | 0.00 | 37.20 |
| 04 - Earth sciences | 39.45 | 1.38 | 26.15 |
| 05 - Biology | 18.75 | 11.31 | 25.89 |
| 06 - Medicine | 6.64 | 4.72 | 15.73 |
| 07 - Agricult. and vet. sciences | 28.43 | 8.48 | 12.22 |
| 10 - Antiq., philol., lit. studies, art hist. | 21.28 | 1.41 | 5.01 |
| 11 - Hist., phil., pedagogy and psychol. | 13.57 | 0.44 | 4.81 |
| 12 - Law studies | 36.52 | 1.74 | 7.83 |
| 13 - Economics and statistics | 25.22 | 0.95 | 4.30 |
| 14 - Political and social sciences | 24.16 | 0.18 | 7.28 |
| Total | 21.83 | 2.65 | 9.71 |

Continuations (B) = continuations in a different course of study of the same degrees class at the University of Bari Aldo Moro. Continuations (C) = continuations in a different course of study of another degree class at the University of Bari Aldo Moro

**Table 4** Outcome in the transition between the 1st and the 2nd year of the course for single-cycle master's degree by CUN scientific area (cohort 2015/2016, University of Bari Aldo Moro). *Source*: National Student Registry of MIUR-Cineca

| CUN Area | Dropouts (%) | Contin. (B) (%) | Contin. (C) (%) |
|---|---|---|---|
| 03 - Chemistry | 12.50 | 0.00 | 12.50 |
| 05 - Biology | 16.30 | 6.00 | 32.10 |
| 06 - Medicine | 2.30 | 1.70 | 0.60 |
| 07 - Agricult. and vet. sciences | 20.00 | 0.00 | 13.30 |
| 11 - Hist., phil., pedagogy and psychol. | 1.60 | 0.00 | 3.30 |
| 12 - Law studies | 20.70 | 0.90 | 5.40 |
| Total | 16.36 | 2.24 | 11.59 |

Continuations (B) = continuations in a different course of study of the same degrees class at the University of Bari Aldo Moro. Continuations (C) = continuations in a different course of study of another degree class at the University of Bari Aldo Moro

– Type of high school diploma;
– High school diploma grade (from 60 to 100 points, plus 100 cum laude = 100L)
– Number of UECs (University Educational Credits) achieved during the first year of the course.

The influence of gender on dropout rates is remarkable (Fig. 5). The data for 2015/2016 cohort show that men are more likely to dropout, with a difference of about 5% points compared to women. In particular, for bachelor's degrees we have 25.8% for men compared to 19% for women. For single-cycle master's degrees we have 19.1% for men versus 14.8%

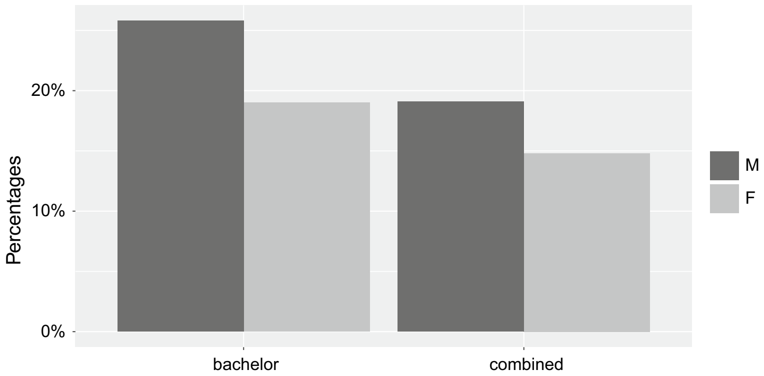


**Fig. 5** Dropout rates between the first and the second year of the course of study, by the type of course and gender (cohort 2015/2016, University of Bari Aldo Moro). *Source*: National Student Registry of MIUR-Cineca

for women. The type of high school diploma also influences the dropout rates (Fig. 6): a high proportion of students from vocational (professional) or technical colleges drop-out university (respectively 26.7% and 35.3%): for students from vocational high schools, dropout rates are 35% for bachelor's and 48% for single-cycle master's degree, respectively. For students from technical high schools, dropout rates are 26.7% (bachelor) and 35.3% (combined).

Even more apparent is the link between the diploma grade and the dropout rate (Fig. 7): in fact, the lower the diploma grade, the more the total dropout rate increases, from 8% for 100 cum laude to 33% for the grade class 60–69 points. Finally, Table 5 shows that students achieving less than 12 UECs in the first year of the course have a dropout rate of 61.7% (54.1% for single-cycle master's and 63% for bachelor's degrees), while those achieving more than 25 UECs present a very low risk of dropout, around 5%. This data is

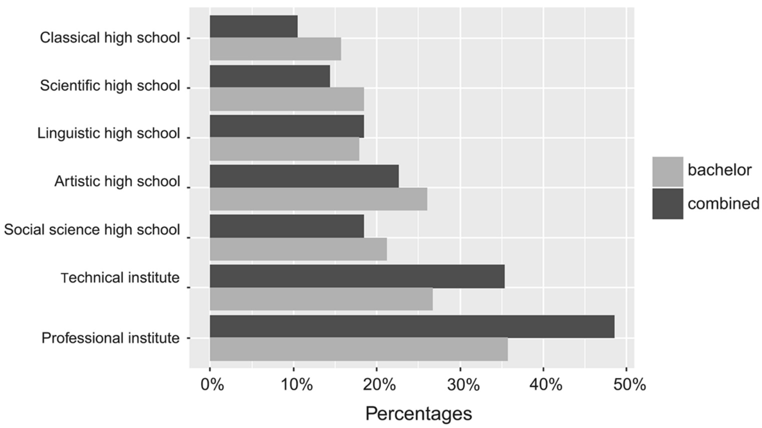


**Fig. 6** Dropout rates between the first and the second year of the course of study, by the type of course and the type of high school diploma (cohort 2015/2016, University of Bari Aldo Moro). *Source*: National Student Registry of MIUR-Cineca

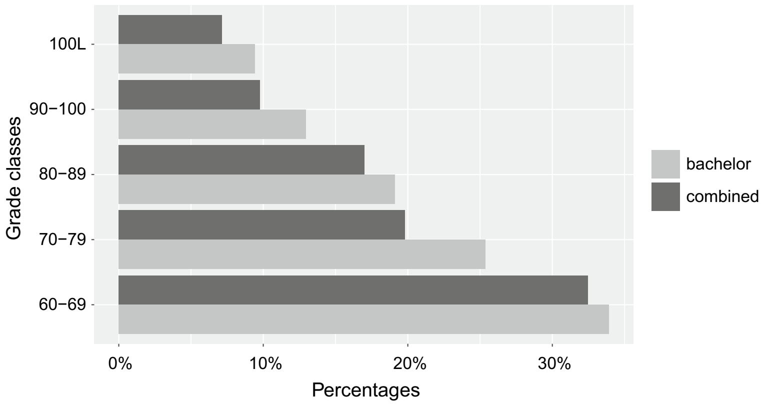


**Fig. 7** Dropout rates between the first and the second year of the course of study, by the type of course and the class of high school diploma grade (cohort 2015/2016, University of Bari). *Source*: National Student Registry of MIUR-Cineca

**Table 5** Dropout rates between the first and the second year, according to UECs achieved in the first year of the course (cohort 2015/2016, University of Bari). *Source*: National Student Registry of MIUR-Cineca

| UECs | Dropouts (%) |
|---|---|
| 0–12 | 61.75 |
| 13–24 | 16.18 |
| 25–36 | 5.86 |
| 36–48 | 1.36 |
| 49–60 | 0.35 |
| > 60 | 0.45 |
| Total | 20.81 |

very important, and will also play a key role in the individual analysis that we will carry out in the next section.

# 4 The Profiles of Students Who Dropout

## 4.1 Education Data Mining and Dropout Prediction

The descriptive analysis that we have carried out so far allows us to highlight the underlying trends and patterns of the phenomenon we are studying, as some variables are strictly correlated with the dropout rates. This has undoubtedly its own importance to facilitate the identification of suitably policy guidelines to reduce dropout rates in future cohorts. However, another specific way of approaching the problem of dropout is an integral part of a broader research field that has emerged over recent years, called 'Educational Data Mining'. The goal is to predict, in an empirical way, the students who are at risk of dropout using a set of input variables associated with the response variable (presence/absence of dropout) or, at least, to identify those variables having a relevant weight in explaining the risk of abandonment at the individual level.

Supervised classification algorithms are well suited to this task (Hastie et al. 2009; Loog 2018). They are based on the availability of a training set, with complete information, in which for each example (instance) of the problem both the classification label (usually it is a 0/1 binary label) and a set of values of qualitative/quantitative input variables are available. Based on this data collection, the algorithm creates an empirical relationship between the space of the input variables and the label, thus making it possible to predict the label also for new future instances, for which only the input variables are available, while the label must still be observed. With an appropriate coding, we can insert the occurrence/non-occurrence of dropout on an individual level within a classification algorithm.

When the goal is to provide a pure decision support system that allows early identification of the students most at risk, in order to be able to implement timely corrective measures that can help reduce the phenomenon in question, there are obvious constraints on the predictors that can be used, in the sense that we are forced to use a minimal set of input variables: by 'minimal' we mean a small set of individual variables that are immediately available at the time of matriculation and that remain constant throughout the university career (not requiring prospective collection of new data). If we do not take these obvious considerations into account, we will systematically obtain results that are not reproducible and are overly optimistic in terms of predictive accuracy (see, for example, the discussion in Márquez-Vera et al. 2016). However, if the objective is to identify the profile of students at higher risk of dropping out, these constraints can be relaxed. In this case, it is more important to explain the risk of dropping out than to predict individual events, and the classification model allows to discover which variables can reproduce, retrospectively and in the best possible way, the dropouts that occurred during the observation period.

Any such out-of-sample predictive analysis can be accompanied by an in-sample analysis (i.e. conditionally to the particular sample that has been observed), in order to get an initial idea of the most relevant variables. In-sample analyses do not have classifying future students as a primary objective, but rather of identifying those variables which, once suitably segmented by means of a classification model, make it possible to identify (with high sensitivity and specificity) students who dropout their courses of study. Therefore, if we are studying the problem of university dropouts retrospectively, we have two points of view that complement each other.

### 4.2 Profiling the Risk of Dropping Out

On the basis of the above discussion, we start the analysis from an exploratory in-sample analysis, using a simple logistic regression model. In what follows, let $c$ denote a binary label with $c \in \{0, 1\}$, the positive class $c = 1$ indicating a dropout. Thus, the response variable $\mathsf{DROPOUT}$ is of dichotomous type, equal to 1 if the student has dropped out and to 0 if not. For each instance (student), a feature vector $\mathbf{x}^\top = (x_1, \ldots, x_s) \in \mathcal{X} \subseteq \mathbb{R}^s$ of $s$ input variables is also available. The regressors introduced in the model are those reported below in Table 6.

The $\mathsf{AREA}$ variable distinguishes the medical-scientific area courses from courses concerning the social-humanistic area, $\mathsf{DIPLOMA}$ the type of diploma achieved by the student suitably dichotomized, the $\mathsf{DEGREE}$ variable distinguishes the type of course of study undertaken (bachelor's or single-cycle master's degree). The $\mathsf{CREDITS}$ variable measures the University Educational Credits achieved by students during the first academic year (2015/2016), while $\mathsf{AGE}$ is the age at enrollment. The high-school diploma grade

**Table 6** Input variables used in the exploratory logistic regression model

| Variable | Coding |
|---|---|
| AREA | 1 = medical/scientific;0 = social/humanistic |
| AGE | Age in years at enrollment |
| GENDER | 1 = M;0 = F |
| DIPLOMA | 1 = classical/scientific high school;0 = other high schools |
| DEGREE | 1 = bachelor's degree;0 = single-cycle master's degree |
| DIPLOMA GRADE | High school diploma grade in cents |
| CREDITS | UECs (University Educational Credits) achieved in 2015/2016 |

(DIPLOMA GRADE) has been introduced into the model because of its high correlation with the risk of dropout (see Fig. 7). The model to be estimated is thus the following:

$$\begin{aligned} \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) =& \beta_0 + \beta_1 \text{ AREA}_i + \beta_2 \text{ AGE}_i + \beta_3 \text{ GENDER}_i \\ &+ \beta_4 \text{ DIPLOMA}_i + \beta_5 \text{ DEGREE}_i \\ &+ \beta_6 \text{ DIPLOMA GRADE}_i + \beta_7 \text{ CREDITS}_i, \end{aligned} \tag{1}$$

where $\pi_i = \Pr\left\{\text{DROPOUT}_i = 1|\mathbf{x}_i\right\}$ indicates the probability of dropout, and $\mathbf{x}_i$ is the vector of values assumed by regressors for student $i = 1, \dots, N$, with $N = 7304$ students in the training sample used here (2015/2016 cohort). R 3.6.0 was used to estimate the model parameters (R Core Team 2019); the results are presented in Table 7.

From both odds ratio (OR) magnitudes and $p$-values it emerges how the following variables have a very significant influence on the risk of dropout ($p < 0.001$, in order of importance): GENDER, CREDITS and AREA. For example, men have a probability of dropout that is 40% higher than that observed in women, while students enrolled in courses that fall into the scientific/medical area have a probability 32% lower than students enrolled in courses of the social/humanistic area. We also observe a less explained effect, that is the probability of dropping out increases as the diploma grade increases, $p < 0.01$ with adjusted OR = 1.010, that is the probability of dropping out increases by 1% for each additional point obtained. It is easily noted that this conclusion is contraddicted by the

**Table 7** Odds ratio (OR) estimates with associated 95% confidence intervals for the exploratory logistic regression model *** $P < 0.001$, ** P < 0.01, and * that P < 0.05).

| Covariate | Est. | se | adj. OR (95% ci) | $p$-value | Signif. |
|---|---|---|---|---|---|
| (Intercept) | − 0.500 | 0.407 | | | |
| AREA | − 0.384 | 0.083 | 0.681 (0.578, 0.802) | < 0.001 | *** |
| AGE | 0.031 | 0.012 | 1.031 (1.007, 1.056) | < 0.05 | * |
| GENDER | 0.338 | 0.081 | 1.402 (1.196, 1.643) | < 0.001 | *** |
| DIPLOMA | − 0.217 | 0.082 | 0.805 (0.685, 0.945) | < 0.01 | ** |
| DEGREE | 0.116 | 0.108 | 1.123 (0.909, 1.386) | > 0.10 | |
| DIPLOMA GRADE | 0.010 | 0.004 | 1.010 (1.003, 1.017) | < 0.01 | ** |
| CREDITS | − 0.127 | 0.003 | 0.881 (0.875, 0.887) | < 0.001 | *** |

aggregated data shown in Fig. 7. However, the crude OR is 0.95, showing that the unadjusted effect of the diploma grade on the probability of dropping out is absorbed into the other input variables when a multivariate model is considered. In this sense the adjusted OR captures a 'pure' effect, that could be adequately explained by the reasons proposed in Belloc et al. (2010) . In the same way, those who attended a classical/scientific high school have a probability of dropout about 20% lower than those who attended other high schools. The contribution of the explanatory variable AGE is statistically significant albeit less strong ($p < 0.05$ with adjusted OR = 1.031, the risk of dropout increases by 3% per additional year, in contrast to Belloc et al. 2010), while DEGREE is not significant.

Apart from GENDER, the most important covariate is CREDITS, with adjusted OR = 0.881, that is the probability of dropouts decreases of 12% for each additional UEC earned during the first year. In other words, students who pass exams during their first year of enrollment have a very slight probability to abandon their studies, while inactive students have a consistent risk of dropping out. Other authors showed the probability of dropout decreases as the academic performance during the first year increases, and therefore the perceived self-regulatory efficacy increases (see, for example: Georg 2009 and Belloc et al. 2011). Overall, the results obtained on the examined collective appear to be in line with those of other similar studies (Chiandotto and Giusti 2005). We point out that the covariate CREDITS can be measured only a posteriori (at the end of the first year): however, as we noted before, our purpose is not to build a pure predictive system that allows early identification of the students at risk, but rather to identify retrospectively the determinants of the phenomenon of the dropouts.

### 4.3 Out-of-Sample Analysis

We now want to analyze the impact of input variables from a predictive point of view, to complement the in-sample retrospective analysis based on logistic regression. The estimation of a decision tree is one of most common statistical techniques used in the literature on the dropout risk (Kingsford and Salzberg 2008; Dekker et al. 2009; Kumar and Pal 2011). Unlike other decision-making models, the decision tree makes all possible alternatives explicit in a transparent way and traces each alternative to its conclusion in a single view, allowing for easy comparisons. However, the determination of the optimal model is not an easy task, as a very large tree might overfit the data, while a small tree could be unable to capture important structures. The preferred strategy is to grow a large tree, stopping the splitting process only when some minimum node size is reached, and then this large tree is pruned using cost-complexity pruning (Hastie et al. 2009). In cost-complexity pruning we define the total cost of a tree $T$ as:

$$C_\alpha(T) = R(T) + \alpha|T| \tag{2}$$

where $R(T)$ is the training misclassification rate, and $\alpha|T|$ is a penalty, where $\alpha \in [0, +\infty[$ is the complexity parameter and $|T|$ is the size of the set of leaf nodes of $T$. When the number of leaf nodes increases with one (one additional split), then the total cost increases with $\alpha$ if $R(T)$ remains unchanged. Depending on the value of $\alpha$, a highly complex tree that makes no errors on the training set may have a higher total cost than a small tree that makes a certain number of errors (on the training set). Under weak technical conditions, given a sequence of complexity parameters $(\alpha_0, \alpha_1, \dots, \alpha_{K-1}, \alpha_K)$ with $\alpha_0 = 0$ and $\alpha_K = +\infty$, it can be shown that it is always possible to construct a sequence of subtrees $T_1 > T_2 > \cdots > T_K$, where $T_k$ is the smallest cost minimizing subtree for any $\alpha \in [\alpha_{k-1}, \alpha_k)$ and $k \in \{1, \dots, K\}$ (Breiman et al. 1984).

The most obvious way to select the final tree from the sequence created with cost complexity pruning is to pick the one with the lowest error rate on a test set or, even better, to use cross-validation (CV) to avoid setting aside a subset of the data for testing. Following the latter approach, we estimated the accuracy on the training set by resampling, using a tenfold CV. In particular, we calculated (taking the average of the values obtained in each of the tenfolds of the training set used as a test set during the CV procedure) the Area Under the Curve (Auc) associated with the ROC curve (Fawcett 2006), as well as sensitivity and specificity (Parikh et al. 2008; Liu 2011):

$$\mathsf{Sens} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}} \tag{3}$$

$$\mathsf{Spec} = \frac{\mathsf{TN}}{\mathsf{TN} + \mathsf{FP}} \tag{4}$$

where TP = True Positives (i.e. the number of actual students dropping out the course of study undertaken that are correctly identified as such) and, obviously, FN = False Negatives, TN = True Negatives e FP = False Positives.

Sensitivity measures the fraction of students, among dropouts, correctly identified by the algorithm. On the other hand, the specificity measures the fraction of students, among all those who have achieved the qualification, which are correctly classified by the algorithm. Optimizing for sensitivity or specificity obviously means pursuing different objectives, and there is a trade-off between the two measures, in the sense that optimizing for one of the two generally means reducing the value of the other. However, greater sensitivity is obviously the most important goal to achieve, since greater sensitivity corresponds to a greater ability to correctly identify the students who leave. Using the infrastructure provided by the R `caret` package (v. 6.0-84; Kuhn 2008), we made the complexity parameter $\alpha$ vary in a suitable way, and we chose the final model as the one which had the highest sensitivity (calculated on the training set by CV in the way described above).

The results obtained are shown in Table 8. Since higher $\alpha$ values correspond to less complex trees, with the same specificity we choose the tree that has the highest value of $\alpha$ (based on an obvious principle of parsimony). The optimal value of $\alpha$ is indicated in bold in Table 8: it corresponds to a sensitivity of 81% (i.e. about eight students out of ten of those who leave are correctly classified) and a specificity of 88% (i.e. almost nine students out of ten of those who graduate are correctly classified). Furthermore, the AUC of the optimal classifier is equal to 0.8440: taking into account the relative standard deviation (reported in column AUCsd) it is evident that the approximate 95% confidence interval for the AUC shows that the classifier obtained has a significantly higher performance than that of the purely random classifier (for which AUC = 0.50). Also the cross-validated accuracy (not shown in Table 8) is maintained at high levels, and precisely it was found to be equal to 86.22%.

Using the entire training set, we also built the tree corresponding to the optimal value of $\alpha$ (Fig. 8). The result obtained is surprising, since the tree obtained is extremely pruned: no variable enters the tree structure except CREDITS. Although it is not a real predictive analysis for the reasons we have already analyzed, but rather a retrospective analysis, the CREDITS variable is able to identify at least eight out of ten among the students who will dropout during the first year. Furthermore, the intrinsic structure of the classification tree,

**Table 8** Grid search of the optimal value of the complexity parameter $\alpha$

| $\alpha$ | AUC | Sens | Spec | AUCsd | Senssd | Specsd |
|---|---|---|---|---|---|---|
| 0.0000 | 0.8978 | 0.6382 | 0.9174 | 0.0161 | 0.0407 | 0.0140 |
| 0.0178 | 0.8593 | 0.6231 | 0.9419 | 0.0173 | 0.0397 | 0.0098 |
| 0.0355 | 0.8593 | 0.6231 | 0.9419 | 0.0173 | 0.0397 | 0.0098 |
| 0.0533 | 0.8588 | 0.6263 | 0.9401 | 0.0174 | 0.0485 | 0.0151 |
| 0.0710 | 0.8446 | 0.7967 | 0.8790 | 0.0164 | 0.0690 | 0.0202 |
| 0.0888 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1065 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1243 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1420 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1598 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1775 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.1953 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.2130 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.2308 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.2485 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.2663 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.2840 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| 0.3018 | 0.8440 | 0.8129 | 0.8751 | 0.0162 | 0.0296 | 0.0123 |
| **0.3195** | **0.8440** | **0.8129** | **0.8751** | **0.0162** | **0.0296** | **0.0123** |
| 0.3373 | 0.7114 | 0.5053 | 0.9176 | 0.1631 | 0.3898 | 0.0640 |

The optimal value (indicated in bold), was obtained by optimizing for sensitivity. For each $\alpha$, sensitivity, specificity and AUC were calculated using tenfold CV, taking the average of all the values obtained on each fold taken as a test set. The AUCsd, Senssd and Specsd columns indicate the standard deviations of the respective indices, also calculated by CV

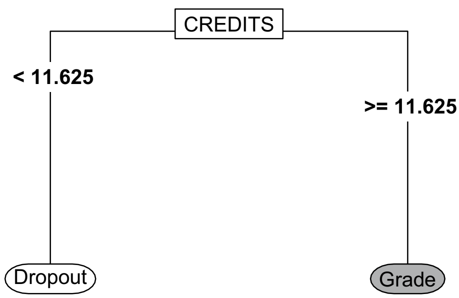


**Fig. 8** Classification tree corresponding to the optimal value of the complexity parameter $\alpha$, built using the entire training set, and optimizing over sensitivity by tenfold cross-validation

characterized by a sequential set of decision rules that partition the input space, leads us to an even more interesting conclusion. The most discriminating threshold (in terms of predictive accuracy) to distinguish students who graduate from those who leave corresponds to (approximately) 12 UECs: in other words, those who pass at least two 6 UEC exams, or at least one fundamental exam with at least 8 UECs plus a 6 UEC exam, have a posterior probability of continuing the studies higher than the posterior probability of dropping out.

This result, in addition to completing the description of the student's leaving profile (based on what has already been achieved through in-sample logistic regression analysis), once again highlights the importance of two fundamental factors: (i) The decision to give up is decisively influenced by the self-perceived effectiveness of the study. If during the year the student manages to pass only a few exams, there will be a greater chance she/he will dropout. These conclusions have been repeatedly reached by other authors (see, in particular, Burgalassi et al. 2016). (ii) The presence of university tutors becomes fundamental to support those students who have difficulties in passing exams. Given the scarcity of resources available for this type of service, tutoring must be fed by informations coming from a true forecasting system, which, on the basis of a set of socio-demographic and performance variables already available at the time of enrollment, allow flagging students who appear to be most at risk of dropping out.

### 4.4 Measuring Variable Importance

The final tree illustrated in Fig. 8 is obtained through a recursive partition of the space of input variables. The tree reaches its maximum size inserting the abovementioned variables according to a measure of the quality of the partition obtained (such as the entropy or the Gini index), and is subsequently reduced through a cost complexity pruning. However, decision trees are notorious unstable, since having a high prediction variance, and small variations in the input data can lead to drastic changes at the end of the procedure. Thus, the result obtained must be confirmed using an alternative analysis technique that has less sensitivity to variations in input data. The standard technique to reduce the prediction variance is known as bagging (Dietterich 2000), and consists in learning $M$ different trees on $M$ randomly chosen subsets of the data. If $\hat{\gamma}_m(\mathbf{x})$ is the empirical classifier corresponding to the tree learned on the $m$-th dataset, for $m = 1, 2, \dots, M$, the final empirical classifier $\hat{\gamma}(\mathbf{x})$ is obtained taking the most frequent label among those assigned by each tree $\hat{\gamma}_m(\mathbf{x})$ (i.e. of the two possible labels, the one that has been assigned most frequently by each of empirical classifiers $\hat{\gamma}_m(\mathbf{x})$).

Proceeding as described above, each of the $M$ subsets is a subset of same data set and, therefore, a certain input variable in one of these $M$ subsets will tend to be strongly correlated with the same input variable in each of the remaining $M - 1$ subsets. So, the reduction in prediction variance will tend to be decidedly limited, since each $\hat{\gamma}_m(\mathbf{x})$ will tend to be very similar to the remaining empirical classifiers. The technique known as random forests (RF) attempts to decorrelate the empirical classifiers taking not only a randomly chosen subset from all the data, but also a random subset of all the input variables (Breiman 2001). For historical reasons, the number of input variables included in the specific training set used by $\hat{\gamma}_m(\mathbf{x})$ is indicated as `mtry`, with $\texttt{mtry} \leq s$. Several papers present in literature have demonstrated the excellent predictive accuracy of this learning technique, explaining its use in diverse applications (Caruana and Niculescu-Mizil 2006).

Nevertheless, our objective is not to use a random forest to improve predictive accuracy, but rather to calculate the importance of each variable in terms of its impact on the predictive accuracy. For each random forest there is a natural way of calculate such impact. First of all, the prediction accuracy on the test sample is measured. Thereafter, the values of a given input variable in the test sample are randomly shuffled, keeping all other variables the same. The accuracy is remeasured after permuting the chosen predictor variable. The difference between the two accuracies are then averaged over all trees, and normalized by the standard deviation of these differences. Features which produce large values for this score are ranked as more important than features which produce small values.

To implement the method we have used a one-hot encoding of input variables (rather than a dummy coding), in order to get exactly one coded variable for each level of the categorical input variables (as it is known that the reference level disappears from the coded data matrix when a dummy coding is used). Moreover, we have further disaggregated the variable AREA into four distinct sublevels (instead of two used with logistic regression), i.e. medical/scientific/social/humanistic, in order to be able to analyze separately the impact of each of the four areas. Using a one-hot encoding the dimension of the space of input variables rose to $s = 13$.

For the choice of final model we have set $M = 500$ and have varied `mtry` between 2 and 9. The optimal final value of the number of variables, chosen by a tenfold cross-validation and optimizing with respect to sensitivity, was $\mathtt{mtry}_{\mathrm{opt}} = 5$. Thereafter, we have repeated the learning procedure training on the full data, and using the optimal value of `mtry` determined before. Lastly, we have calculated the importance of each variable directly on the full data set (as a training/test splitting was unavailable in our case). This is not to be considered limitative, even though the predictive accuracy estimated on training test is generally a far too optimistic estimate of true accuracy. However, in our case we are interested in a difference of accuracies (calculated before and after the reshuffling of the variable whose importance is being measured). If both accuracies are biased roughly to the same amount, their difference will be approximately unbiased. Therefore, the use of a separate test set is not essential to estimate the importance of a given input variable. The results are shown in Fig. 9, where the most important 10 input variables are reported. By convention, a value of 100 has been attributed to the most important variable (i.e. having an importance equal to 100%).

As may be noted, the most important variable is CREDIT. So, this additional analysis fully confirms the results presented above. It should also be noted how the final mark of a diploma has an importance of roughly at 25%, and therefore not negligible. Furthermore, AGE has an importance equal approximately to 12% of CREDIT. However, further analysis on new data will be necessary, because the direction of the association between the success in studies and the age at enrollment is not clear. Some studies indicate a positive correlation (for example, Belloc et al. 2010), i.e. students who enroll late have a great motivation to complete their studies, while other studies (as ours) indicate a negative correlation, in

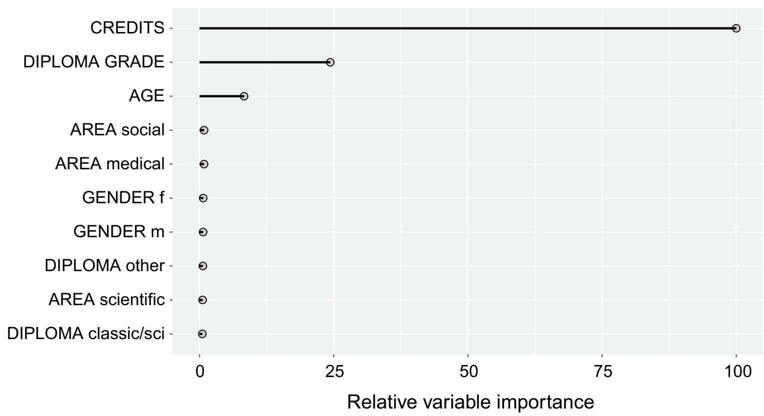


**Fig. 9** Relative importance of input variables. The importance was calculated using a random forest classifier. See the text for details

the sense that it becomes even more difficult to graduate as age advances. The remaining variables are absolutely uninfluential.

## 5 Final Remarks

Dropouts have long been indicated as one of the main pathologies of the Italian higher education system. The objective of reducing the size of the phenomenon and its negative impact on the productivity of the system and on the profitability of the investment in education by the public sector and private individuals (students and their families) is one of the qualifying elements of the reform of the educational offer and, more generally, of system reforms carried out in Italy since the 2000s. The empirical evidence shows that the main motivation for dropping out is the difficulty in passing the exams, with the consequent fall in individual motivation and loss of confidence in personal abilities. Other negative circumstances (that are not investigated here) might be the lack of continuity in the course of studies and the low level of attendance at the university, factors which can reduce the possibility of mobilizing resources and devising strategies to combat difficulties and delays.

A specific analysis carried out on the students of the University of Bari Aldo Moro indicates that the risk of dropping out is greater for inactive (less than 12 UECs achieved) male students, graduated from professional or technical institutes. Preparation gaps, insufficient knowledge of the university environment, poor mastery of effective study methodologies are elements that can negatively affect students' careers. It is advisable to adequately monitor these conditions both on entry and during the course of studies, especially in the initial phase, which is strategic to define the chances of success or failure. The University of Bari has launched numerous initiatives to reduce both the dropout rate and transfer applications in the transition between the first and second year. These initiatives are concerned with both new student orientation, as well as with students of the 4th and 5th years of secondary school (Open Day, Orientation Week, etc.), and with students of the first year of the course, supported by ongoing tutoring activities (each student has its own teaching tutor).

However, the difficult conditions of the labor market may also have a direct negative influence on enrollment and continuation of studies, particularly on young people facing difficult individual or family economic conditions, which can make their educational objectives difficult to reach. Ultimately, therefore, guidance, counseling and support on matriculation for new students appear essential, as well as accompanying and support interventions during the studies, by means of tutoring services and other tailor-made interventions aimed at reducing the dropout rate. Moreover, financial support (scholarships and accommodation for students in poor economic conditions) is likely to be necessary to minimize the total number of higher education dropouts. These aspects will be subject to future research.

Finally, as we said before, setting up a true forecasting system is crucial to allow flagging students who appear most at risk to dropping out. One of the areas in which EDM can play an important role is precisely the early identification of students who are at risk of leaving university studies (Delen 2010; Hoffait and Schyns 2017). The use of artificial intelligence and Machine Learning algorithms (ML) has caused a real paradigm shift in statistical science over the last 10 years (Dunson 2018), which could essentially contribute to develop information systems suitable to this purpose. For example, feedforward networks with a large number of hidden levels, or networks with more complex topologies, but equally characterized by the presence of a very large number of compositions of non-linear functions to model the relationship between input and output, have a higher (and substantially not yet explained) generalization capability than traditional algorithms (LeCun et al. 2015; Kawaguchi et al. 2017). The use of deep learning algorithms, in conjunction with the availability of an adequate amount of information, could therefore lead to a significant performance boost in terms of predictive accuracy and could represent a decisive step forward in building systems of early dropout prediction that can also be used from a practical point of view. These aspects will also be subject to future experimentation and research.

**Acknowledgements** We would like to thank the Rector of the University of Bari for authorizing the consultation, in an anonymous form and for scientific research purposes, of the data relating to the UniBA student population of the MIUR-Cineca Student-Didactic Observatory.

**Authorship contribution** Abbreviations: Paola Perchinunno (PP), Massimo Bilancia (MB), Domenico Vitale (DV). PP and MB conceived the study; DV contributed to the study design. PP, MB and DV wrote the first draft of the manuscript. All authors equally contributed to the writing of Sect. 5. PP wrote Sects. 1, 3.2 and 4.2; MB wrote Sects. 3.1, 4.1 and 4.3, and cared about the overall paper structure; DV wrote Sects. 2 and 4.4.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References


Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*, *53*(3), 563–574.

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

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

Belloc, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: A semiparametric mixed-effects model for an Italian case study. *Journal of Applied Statistics*, *38*(10), 2225–2239.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth, NY: Chapman and Hall.

Burgalassi, M., Biasi, V., Capobianco, R., & Moretti, G. (2016). The phenomenon of early college leavers. A case study on the graduate programs of the Department of Education of "Roma Tre" University. *Italian Journal of Educational Research*, *Anno IX (17)*.

Carci, G. (2018). Gl Immatricolati e gli Iscritti. In *ANVUR national agency for the evaluation of universities and research institutes: Rapporto Biennale sullo Stato del Sistema Universitario e della Ricerca 2018, ANVUR*, chap I.1.3, (pp. 43–73).

Carletti, V., & Infurna, M. R. (2018). I Percorsi di Studio: Mobilità, Abbandoni e Conseguimento del Titolo. In *ANVUR national agency for the evaluation of universities and research institutes: Rapporto Biennale sullo Stato del Sistema Universitario e della Ricerca 2018, ANVUR*, chap I.1.3, (pp. 43–73).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning—ICML '06* (pp. 161–168). New York, USA: ACM Press.

Chiandotto, B., & Giusti, C. (2005). L'abbandono degli studi universitari. In C. Crocetta (Ed.), *Modelli Statistici per l'Analisi della Transizione Università-Lavoro* (pp. 1–2). Padova: CLEUP.

Cipollone, P., & Cingano, F. (2007). University drop-out—-The case of Italy. Bank of Italy Temi di Discussione (Working Paper) No. 626.

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). Predicting students drop out: A case study. In T. Barnes, M. C. Desmarais, C. Romero, & S. Ventura (Eds.), *Second International Conference on Educational Data Mining (EDM 2009) July 1–3, 2009* (pp. 41–50). Spain: Cordoba.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.
Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple classifier systems* (pp. 1–15). Berlin: Springer.
Di Pietro, G. (2004). The determinants of university dropout in Italy: A bivariate probability model with sample selection. *Applied Economic Letters*, *11*(3), 187–191.
Di Pietro, G., & Cutillo, A. (2008). Degree flexibility and university drop-out: The Italian experience. *Economics of Education Review*, *27*(5), 546–555.
Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, *136*, 4–9.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.
Georg, W. (2009). Individual and institutional factors in the tendency to drop out of higher education: A multilevel analysis using data from the Konstanz Student Survey. *Studies in Higher Education*, *34*(6), 647–661.
Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York, NY: Springer.
Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11.
Ismail, M. R., Awang, M. K., Rahman, M. N. A., & Makhtar, M. (2015). A multi-layer perceptron approach for customer churn prediction. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(7), 213–222.
Johnson, J. L. (1997). Commuter college students: What factors determine who will persist or who will drop out? *College Student Journal*, *31*(3), 323–332.
Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2017). Generalization in deep learning. arXiv:1710.05468v4 [stat.ML].
Khodabandehlou, S., & Zivari Rahman, M. (2017). Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. *Journal of Systems and Information Technology*, *19*(1/2), 65–93.
Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology*, *26*(9), 1011–1013.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.
Kumar, B., & Pal, S. (2011). Mining educational data to analyze students performance. *International Journal of Advanced Computer Science and Applications*, *2*(6), 63–69.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Liu, B. (2011). *Web Data mining: Exploring hyperlinks, contents, and usage data* (2nd ed.)., Data-centric systems and applications Berlin: Springer.
Loog, M. (2018). Supervised classification: Quite a brief overview, chapter 5. In E. Camporeale, S. Wing, & J. R. Johnson (Eds.), *Machine learning techinques for space weather* (pp. 113–145). New York: Elsevier.
Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, *33*(1), 107–124.
Meggiolaro, S., Giraldo, A., & Clerici, R. (2017). A multilevel competing risks model for analysis of university students' careers in Italy. *Studies in Higher Education*, *42*(7), 1259–1274.
Miguéis, V., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36–51.
Mitchell, T. M. (1997). *Machine learning* (1st ed.). New York, NY: McGraw-Hill Inc.
Montmarquette, C., Mahseredjian, S., & Houle, R. (2001). The determinants of university dropouts: A bivariate probability model with sample selection. *Economics of Education Review*, *20*(5), 475–484.
Murray, M. (2014). Factors affecting graduation and student dropout rates at the University of KwaZulu-Natal. *South African Journal of Science*, *110*(11/12), 1–6.
Organization for Economic Co-operation and Development. (2017). OECD skills strategy diagnostic report: Italy 2017, OECD 2017.
Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50.
Paura, L., & Arhipova, I. (2014). Cause analysis of students' dropout rate in higher education study program. *Procedia - Social and Behavioral Sciences*, *109*, 1282–1286.

R Core Team. (2019). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Smith, J. P., & Naylor, R. A. (2001). Dropping out of university: A statistical analysis of the probability of withdrawal for UK university students. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, *164*(2), 389–405.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.