Investigating Student Retention in an Economics Degree Programme

Francesca Pierri[1(✉)] and Chrys Caroni[2]

[1] Department of Economics, University of Perugia, Perugia, Italy
francesca.pierri@unipg.it
[2] Department of Mathematics, National Technical University of Athens, Athens, Greece
ccar@math.ntua.gr

Abstract. For financial and reputational reasons, every university and degree course is committed to the goal of increasing the number of enrollments and improving indicators of students' success, including decreasing the time taken by students to obtain their degree and reducing the incidence of dropping out. High rates of dropping out from university courses and excessive time to graduation are matters of great practical importance, with substantial financial and social costs. Student retention or dropout depends on a combination of various factors, the influence of which is mediated by the student's ability to integrate into the academic system. The identification of those factors which may affect the outcome and upon which the academic system can act, is an object of extensive study and research.

In order to examine this topic, we collected data at the University of Perugia at the time of initial registration for three bachelor degrees in Economics. The 2015 cohort of students was tracked up to 2025, a period substantially exceeding the nominal duration of their degree courses, allowing the determination of each student's current status at the time of data extraction.

Applying binary and multinomial logistic regression, survival analysis and competing risks (graduation and dropout) analysis, we explore which factors influence the students' performance and highlight pros and cons of the applied statistical methodologies in this context.

Keywords: university dropout · student retention · logistic regression · survival analysis · competing risks

1 Introduction

One of the primary challenges faced by universities in the last decade is to increase the number of students enrolled annually and the expansion of the provision of university training through the opening of new courses has surely encouraged this growth. Concurrently, universities are busy enhancing their support to reduce the extent of the phenomenon of dropout, and improve and accelerate

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15887, pp. 377–392, 2026.
https://doi.org/10.1007/978-3-031-97589-9_26

academic study pathways, thus assisting young people in understanding their aptitudes, capabilities, and desires through guidance programmes that could help students to choose the University path that suits them best. These aims have both social and economic aspects; in Italy, the funds that a University receives from the Ministry of University and Research for ordinary operations (FFO) depend on the number of students regularly enrolled with a period of stay within a bachelor's degree no longer than four years (the nominal length of the degree programme being three years). Simultaneously, the annual enrolment fee paid by students rises with the length of their stay. Regional subsidies, such as those for canteen access or grants, decrease, leading to increased household spending. Recent reviews [1,3] state that student retention depends on a mixture of individual and external factors, which are not always controllable and have determinants not known a priori. Consequently, our study focuses attention on personal characteristics, rather than on others where there is an impossibility of intervention during the academic course of the students: sociodemographic background or external conditions. In this study, we focus attention on students' characteristics, educational background and performance, instead of on behavioural and relational aspects such as integration into academic and social life.

For this purpose, we used data provided by the students at the time of registration and collected in the University's official database. A cohort of students enrolled in 2015 in the three bachelor degrees in Economics taught by the University of Perugia was followed up to 2025. As a first step, in the present paper we aim to use these data to contribute to the identification of the strongest predictors of the risk of dropping out. In a subsequent analysis, additional data on the annual credits obtained and the academic performance of the students will be taken into account. This section is followed by a brief description of the methodology applied and then the Application section, data description and results illustration are provided. Finally, the Conclusion highlights the main findings, differences between the methodologies applied and future research directions.

2 Methodology

An event probability may be explored in a variety of ways, applying different statistical methodologies that are well known in risk analysis. The response variable may be treated as binary or multinomial, depending on the number of possible outcomes that represent objects of study (two or more); moreover we may or may not take into account the time of the occurrence of the outcome event (or occurrences if more than one), and whether these events are in competition with each other.

Here, we applied binary logistic regression to investigate the probability of academic success (that is, graduation) and a multinomial logistic regression to consider different causes of dropping out. Next, a classical survival analysis was applied considering as censored all the observations without a closure date at the time of data extraction (that is, students whose university career was still

incomplete) and a competing risk model to analyse the survival curve for each different event, considering others as competitors. The development of models was preceded by descriptive analysis. SAS® software was used, specifically the procedures LOGISTIC [27] and PHREG [26] for logistic regression and survival analysis, respectively. A brief reminder of logistic and survival models is provided in the following Section.

In addition, a high performance procedure, HPSPLIT [25,28], often employed for tasks such as customer segmentation [11], risk assessment [15,21] and predictive modelling [7,10,14,19,30], was applied. This procedure builds tree-based statistical models for classification and regression for either continuous or categorical variables. In our context, the decision tree method was applied mainly for the advantage it offers in terms of interpretability and visualisation. During the decision tree building process, the Gini index and entropy method are used to assess candidate splits for each node. Moreover, a smaller subtree is obtained, starting from the full tree, applying a cost-complexity pruning.

2.1 Logistic Models

The logistic model, widely applied in various field such as medicine [4], social sciences [20], economics and finance [17], and machine learning [31], is a statistical method that models the probability of a binary outcome based on one or more predictor variables. Let Y be the Bernoulli random variable taking value 1 when the event occurs (award of the university degree) and 0 otherwise (academic course not completed or interrupted). Let $\mathbf{x}' = (x_1, x_2, ..., x_p)$ be a covariate vector representing risk factors thought to be related to the event under study. The logistic model is expressed by

$$P(Y = 1|\mathbf{x}) = \frac{e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}}{1 + e^{\alpha+\boldsymbol{\beta}'\mathbf{x}}} \tag{1}$$

where α is the intercept, and $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown coefficients whose value, if positive, indicates an increase in the likelihood of the event as the value of the corresponding covariate increases, or a decrease, if negative. Results may also be evaluated in terms of the odds ratio (OR), where $\text{OR} = e^{\beta}$, which is the change in the odds of the outcome occurring for a one-unit increase in the value of the predictor variable: when an OR is greater than 1 the event is more likely to happen and the opposite when it is less than 1, while when it is equal to 0 there is no effect.

A multinomial logistic regression model is applied when there are more than two possible outcomes

$$P(Y = j|\mathbf{x}) = \frac{e^{\alpha_j+\beta_{1j}x_1+\beta_{2j}x_2+\ldots+\beta_{pj}x_p}}{1 + \sum_{k=1}^{J-1} e^{\alpha_k+\beta_{1k}x_1+\beta_{2k}x_2+\ldots+\beta_{pk}x_p}} \tag{2}$$

$P(Y = j|\mathbf{x})$ expresses the probability of the outcome j given the predictors $\mathbf{x}$; $\alpha_j, \beta_{1j}, \ldots, \beta_{pj}$ are the coefficients for the predictors for outcome j, with J indicating the number of categories.

In our case the categorical dependent variable is of nominal type: one category is chosen as the reference, and the model estimates the odds of the other categories relative to this reference.

Maximum likelihood estimation was used to estimate the parameters of the logistic regression models, carried out by the LOGISTIC procedure [27] using either the Fisher scoring algorithm or the Newton-Raphson algorithm.

To identify the optimal model, stepwise selection was implemented after evaluating multicollinearity among variables. The significance of the individual predictors was assessed through the Wald test: a $p\text{-}value < 0.25$ was fixed for each parameter to enter into the model and a $p\text{-}value < 0.05$ to stay in. Therefore the final model shows only the predictors that were statistically significant at the fixed $p\text{-}value < 0.05$. The Hosmer and Lemeshow Goodness-of-Fit test [13] was used to assess how well the logistic regression model fits the data.

2.2 Survival Models

Survival models are widely used in risk analysis to analyze the expected duration of time until one or more events occur [6], such as onset of a disease [18] or death in biological organisms [2], failure in mechanical systems [24] or bankruptcy in firms [22,32]. This type of analysis is also known as time-to-event analysis or event history analysis. After setting start and end dates for an observational period, the exit date and corresponding cause are recorded, while subjects still alive (that is, without an event) at the end of observation are considered to have right-censored event times.

Let T be the random time until the event and $\boldsymbol{\beta}$ the vector of coefficients denoting the effects of the vector of covariates $\boldsymbol{x}$ upon T. The values of these covariates for each student are measured at time 0 and are time independent. The survival function

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x})\,, \tag{3}$$

expresses the probability that a subject will survive until at least time t. This function can conveniently be estimated non-parametrically using the Kaplan-Meier estimator and the equality of survival curves that have been obtained for disjoint subgroups of the data can be tested by the logrank test. The cumulative incidence function that gives the probability of the event under study up to time t, is

$$F(t|\mathbf{x}) = P(T \leq t|\mathbf{x}). \tag{4}$$

An important role is played by the hazard function $h(t)$ that expresses the instantaneous event rate at time t among subjects that have survived that long:

$$h(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t | T > t]}{\delta t} = \frac{f(t)}{S(t)}\,. \tag{5}$$

In the popular semi-parametric proportional hazards model (Cox regression [8]), the covariates act multiplicatively on an unspecified baseline hazard rate $h_0(t)$ through the factor $e^{\boldsymbol{\beta}'\mathbf{x}}$:

$$h(t|\mathbf{x}) = h_0(t) \exp\left\{\boldsymbol{\beta}'\mathbf{x}\right\} \tag{6}$$

The effect of covariate x_i is expressed by the hazard ratio e^{β_i}: an increase of one unit in the value x_i (while keeping the values of other covariates constant) multiplies, from (6), the hazard by this amount.

Competing risk survival models are used when individuals in a study can experience multiple types of events, and the occurrence of one event prevents the occurrence of others. Two different approaches exist in this field: the cause-specific hazard model and the subdistributional hazard model or Fine & Gray model [9,12,23]; in the present context we consider the latter, which focuses on the probability that the event of interest occurs over time, accounting for the fact that other competing events might take place.

The subdistribution hazard function estimates the risk of failure from cause k,

$$h_k^s(t) = \lim_{\delta t \to 0} \frac{P[t < T \leq t + \delta t \bigcap D = k | T > t \bigcup (T \leq t \cap D \neq k)]}{\delta t}, \qquad (7)$$

and expresses the instantaneous rate of failure at time t among students that have not yet experienced an event of type k.

3 Application

3.1 Data

The University of Perugia provided data on the 2015 cohort of students who entered three bachelor's degree programmes in Economics: Business Administration in the provinces of Perugia (programme code L021) and Terni (programme code L025) (two different districts of the Umbria Region) and Business for Tourism (programme code L026). This cohort was followed from first registration in 2015 until January 2025. Table 1 shows the distribution of these 515 students by course and gender. For each student we have an anonymous numerical identification code, information about gender, region and district of residence, type of secondary school attended, diploma obtained and the related mark. Moreover, we know the student's current status: that is, graduated (status = 1); officially dropped out from studies, which is often done in order to switch to another degree course (status = 2); deleted, if at least eight years have passed since registration and no examinations have been taken in the last three years or if three years have elapsed since the last payment of tuition fees without passing any examination (status = 3); or active, if at least one examination has been taken even if with negative result (status = 0). The dates of change of status are known. States 1, 2 and 3 each marks the end of the student's career in this university, so they represent competing risks in the terminology of survival analysis. Students in state 0 have not yet reached an endpoint, so they have right-censored durations of studies.

In Table 2, which shows the distribution of students' status in 2025 by course, we see that 54.6% of the students obtained their degree within 10 years, while as many as 37% of the initial entrants had ceased to be students on these courses for various reasons, and 8.2% had not yet reached a conclusion.

Table 1. Distribution of students by degree course and gender

Degree Code	Female		Male		Total	
	N	%	N	%	N	%
L021	150	67.6	236	80.5	386	75.0
L025	43	19.4	40	13.7	83	16.1
L026	29	13.1	17	5.8	46	8.9
Total	222	43.1	293	56.9	515	100

Table 2. Distribution of students' current status by course

Degree Code	Sex	Status				
		0	1	2	3	Total
L021						
	F	8	95	23	24	150
	M	16	120	60	40	236
L025						
	F	8	20	8	7	43
	M	7	24	6	3	40
L026						
	F	1	17	5	6	29
	M	2	5	3	8	17
Total	N	42	281	101	88	515
	%	8.2	54.6	20.2	17.1	100

Table 3. Distribution of type of school attended by gender.

School		Female	Male	Total
Classical	N	45	24	69
	%	20.27	8.19	13.40
Scientific	N	45	113	158
	%	20.27	38.57	30.68
Mixed*	N	21	1	22
	%	9.46	0.34	4.27
Professional	N	20	21	41
	%	9.01	7.17	7.96
Technical	N	86	122	208
	%	38.74	41.64	40.39
Outside Italy	N	5	12	17
	%	2.25	4.1	3.30
Total	N	222	293	515

*Includes elements of classical and scientific

Looking at the students' background (Table 3), the higher percentage held a technical (40.4%) or scientific (30.7%) certificate (or secondary school diploma) with significant differences by gender (χ^2 test of association =54.22 with 5° of freedom, $p < 0.0001$, Cramer's V = 0.32); relatively fewer female than male students had attended a scientific secondary school, while relatively more had attended schools of classical or mixed type.

The mean graduation mark was 80.4. Examination of Q-Q plots indicated that marks were not normally distributed, hence Wilcoxon tests rather than t-tests were used to compare marks between groups of students. A significant difference ($p < 0.001$) was found between females (82.9) and males (78.1). There were no significant differences in this respect between the three bachelor's degree programmes.

Table 4. Effects of covariates in Binary Logistic Regression: Parameter Estimate, Standard Error (SE), Wald χ^2 and its probability, Odds Ratio (OR) and Wald 95% Confidence Interval (CI)

	Estimate	SE	Wald χ^2	Pr>χ^2	OR	95% CI
Intercept	-3.816	0.665	32.950	<.0001		
School type*						
Mixed	-1.350	0.488	7.667	0.006	0.259	0.100 - 0.674
Classical	-0.216	0.311	0.482	0.488	0.806	0.438 - 1.482
Professional	-1.956	0.412	22.550	<.0001	0.141	0.063 - 0.317
Outside Italy	1.771	0.976	3.293	0.070	5.876	0.868 - 39.795
Technical	-0.835	0.234	12.746	<.0001	0.434	0.274 - 0.686
School Grad.mark	0.060	0.009	46.379	<.0001	1.061	1.043 - 1.080

Reference category: *Scientific
Hosmer and Lemeshow Goodness-of-Fit Test $p = 0.26$

3.2 Results

Logistic Regressions. The binary logistic regression model obtained after a stepwise selection procedure (Table 4) identified the school graduation mark and the type of high school attended as significant predictors (Wald test, $p<0.05$) of academic success among those factors that had been recorded at the date of enrolment (sex, graduation mark, high school attendance and residence).

More specifically, while there were no significant differences between students who attended high schools with scientific and classical orientations, these were much more likely to be successful than students who had attended technical (Odds Ratio = 0.434, 95% Confidence Interval 0.24–0.686), professional (OR = 0.141 CI = 0.063−0.317) or other secondary schools (OR = 0.259 CI = 0.100−0.674). Moreover, the higher the graduation mark, the higher was the probability of avoiding drop-out (OR = 1.061 CI = 1.043−1.080).

For further analysis of data and results, justified on the basis of other research [21] the high performance procedure HPSPLIT that creates the decision tree

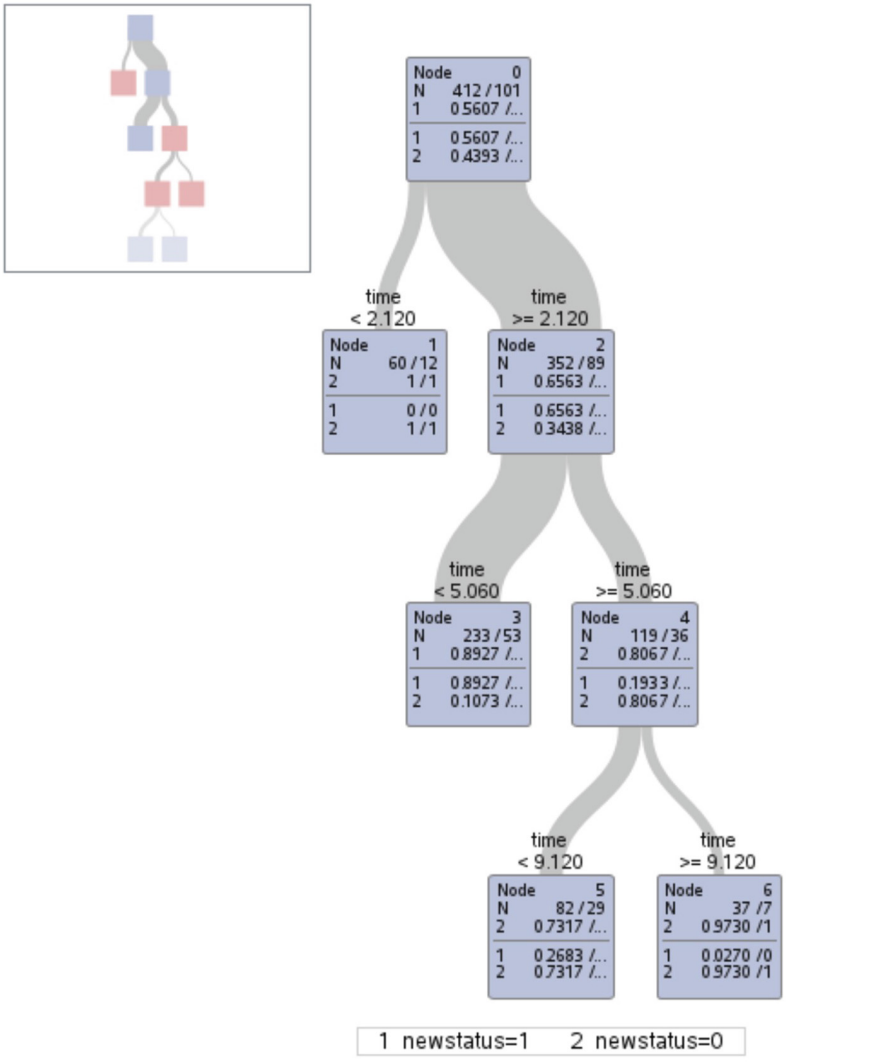

Fig. 1. HPSPLIT output: newstatus = 1 represents the students who graduated, newstatus = 0 is the group with interrupted studies.

model was applied. The length of stay within the degree programme was included among the other covariates in this analysis. As seen in Fig. 1, the first branching is evident at time ≤ 2.120, a time shorter than the duration of any course, in fact, here are allocated all subjects belonging to group 0, while a high percentage of graduated students are assigned to node number 3 when time is less than 5.06. Upon investigation, the group in node number 2 was found to include students who had formally abandoned their studies (status 2), probably in order to

Table 5. Effects of statistically significant covariates ($p<0.05$ for Wald χ^2) in Multinomial Logistic Regression: Odds Ratio (OR) and Wald Confidence Interval (CI)

	status = 0		status = 2		status = 3	
	OR	95% CI	OR	95% CI	OR	95% CI
School type*						
Mixed	0.802	0.091–7.09	4.164	1.298–13.362	8.211	2.338–28.842
Classical	0.702	0.214–2.303	1.411	0.658–3.026	1.582	0.622–4.022
Professional	0.780	0.088–6.932	8.083	2.943–22.202	18.107	6.123–53.553
Outside Italy	0.916	0.071–11.737	0.135	0.013–1.419	0.046	0.004–0.569
Technical	1.549	0.718–3.342	2.126	1.184–3.816	3.484	1.727–7.027
Degree course**						
L025	3.047	1.459–6.364	0.715	0.363–1.408	0.676	0.306–1.494
L026	1.137	0.297–4.355	0.492	0.187–1.298	1.092	0.462–2.582
School Grad. Mark	0.968	0.941–0.996	0.946	0.926–0.967	0.917	0.893–0.942

Reference categories: status = 1, *Scientific, ** L021
Hosmer and Lemeshow Goodness-of-Fit Test $p = 0.86$

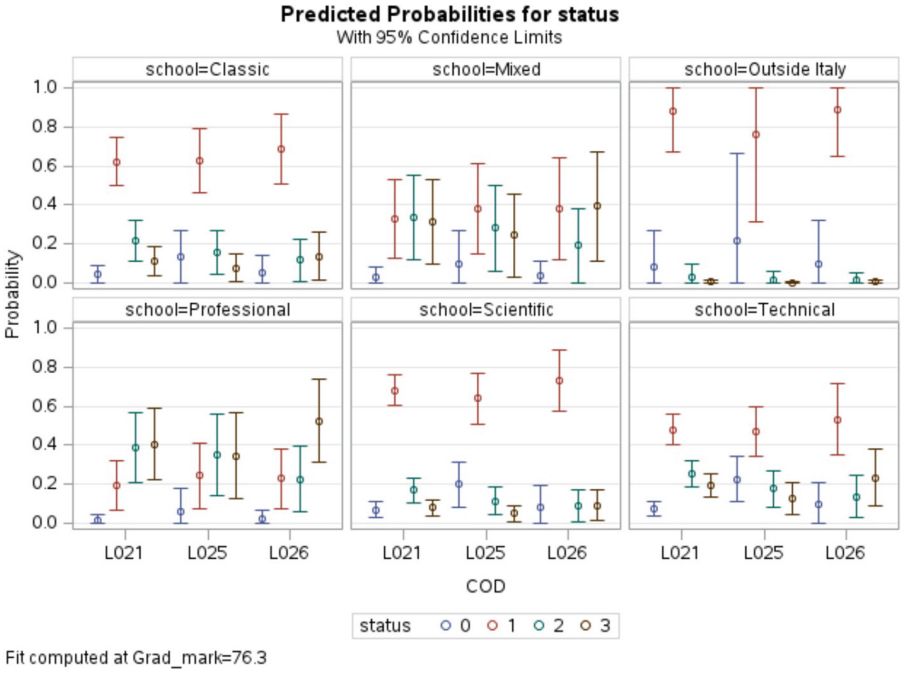

Fig. 2. Multinomial Logistic Regression predicted probabilities: Degree course, and school attended.

switch to another degree programme. Therefore, multinomial logistic modelling was applied, considering the four different ten-year outcomes and using the group of students who successfully completed their studies as the reference. The most significant factors, highlighted after a stepwise selection procedure, and having a Wald χ^2 with $p<0.05$, were the degree course, the school graduation mark and the type of high school attended. (For greater clarity, Table 5 presents the odds ratios and their corresponding confidence intervals for the covariates selected in the stepwise model. The relation between the odds ratio and the parameter estimate is explained in the methodology section). Students enrolled in the L025 course were more likely to extend their academic career beyond eight years (status = 0) in comparison to students enrolled in the L021 course: OR = 3.047 with 95% Confidence Interval 1.459–6.363. On the other hand, these two courses did not differ regarding the other statuses, because the odds ratio includes the number 1. Moreover, (taking the scientific high schools as reference category) a technical or professional secondary school background made it more likely that students would not complete their studies (status = 2), or would drop out (status = 3). For these two statuses, the respective odds are quite high but with wide confidence intervals.

The odds ratios below one for the three different causes of failure indicate the importance of the school graduation mark: as it increases, the probability of dropping out in any way decreases.

Observing Fig. 2 we notice that for a school graduation mark equal to 76.3, the highest probability of success (status = 1) is for students who attended a classical or scientific secondary school or a school outside Italy.

Furthermore, the contribution of the students' graduation mark is clearly shown in the graph obtained as an output from the HPSPLIT procedure applied to assess variable importance in the multinomial model (Fig. 3): the first split is into two branches for graduation marks less than or greater than 76, and subsequently the distinction by type of secondary school attended is evident.

Survival Analysis. The Kaplan-Meier non-parametric estimates of the survival functions by school of attendance (Fig. 4) and degree course (Fig. 5) show that students coming from a scientific or classical school, or enrolled in course L021, have significantly lower probability of remaining within the university for more than three years than students from other schools or from other courses, respectively; that is they have a shorter academic career with a better chance of success within the normal course duration. Similar analyses did not show significant differences by gender or region (Logrank test $p > 0.05$).

The classical survival analysis (Cox regression), after stepwise selection, returned a model with results that are in line with those obtained in binary logistic regression: the significant predictors on the basis of the Wald test were school and graduation mark. In order to test the validity of the proportional hazards assumption for these two predictors, their interaction with time was added to the model [29]. This inclusion turned out to be significant (Wald $\chi^2, p < 0.001$). It appeared that both predictors' influence decreased over time, in particular

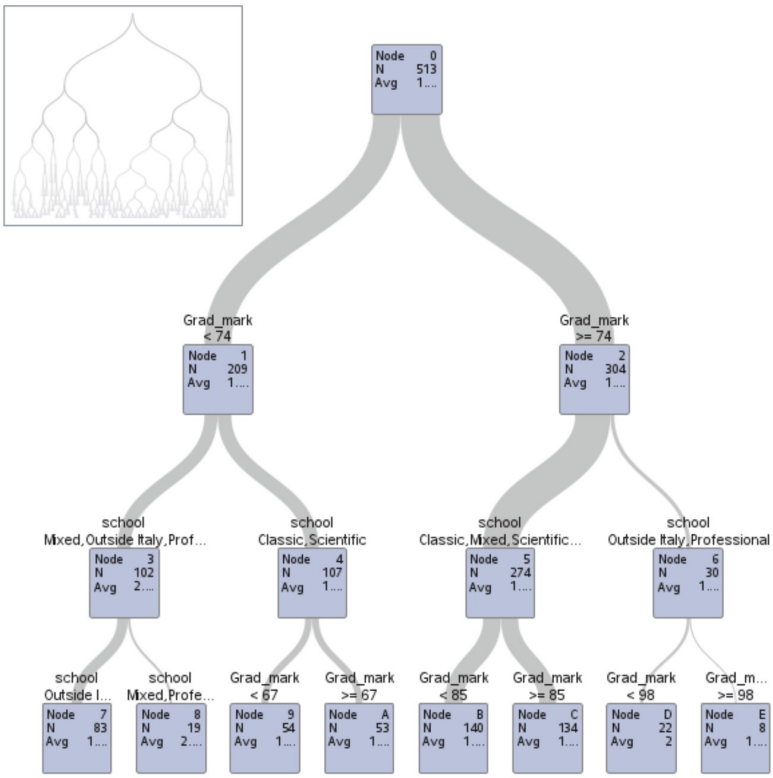

Fig. 3. HPSPLIT tree output: Assessing variable importance.

the graduation mark coefficient has a value equal to 0.223 with a hazard ratio equal to 1.25 meaning that each unit increase in graduation mark increases the hazard by 25%, while the interaction with time, having a negative coefficient value (−0.159) with a hazard ratio less than one shows that the effect of the mark decreases over time.

The Fine and Gray subdistributional hazard model was applied to verify if there were differences by status. In this analysis we considered as censored the students still active at the time of data extraction (status = 0) and we ran three different survival models, one for each status, considering the other two as competing.

The proportional hazard assumption was found to be violated both for the school and graduation mark predictors, but only for events 2 and 3. Therefore interactions with time were included and after this different models were obtained for each event.

For status 1 the graduation mark and the high school attended were again the most significant factors. A hazard ratio greater than one with confidence interval not containing the number 1, fosters belonging to the group. Therefore,

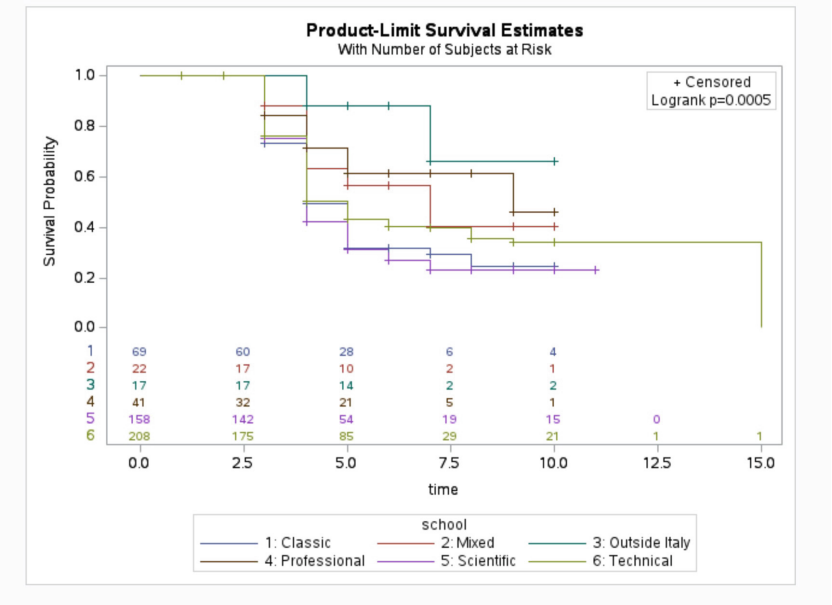

Fig. 4. Kaplan-Meier estimate of the survival curves by type of school attended: here, "survival" means that the student is still active, that is, has not yet graduated.

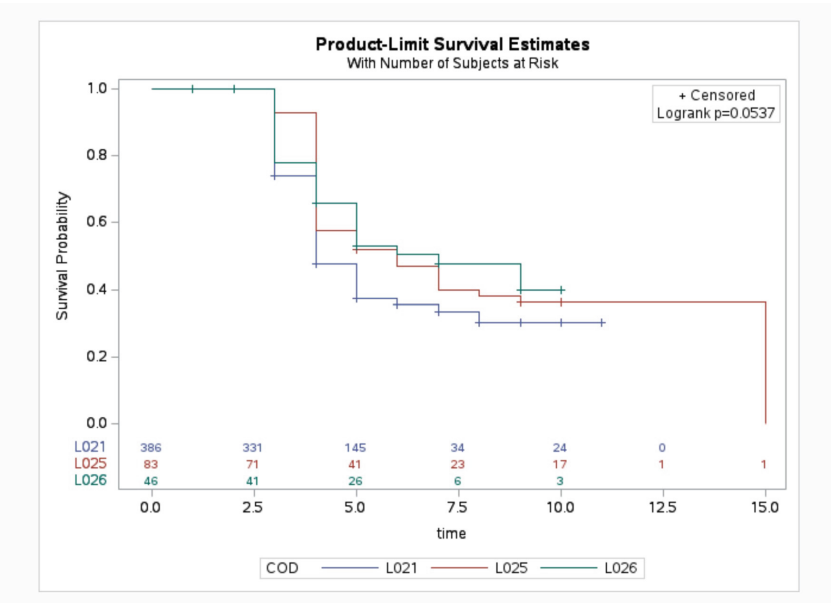

Fig. 5. Kaplan-Meier estimates of the survival curves by university course.

Table 6. Effect of covariates in Fine & Gray subdistributional hazard regression model: Hazard Ratio (HR) and Confidence Interval (CI)

	status = 1		status = 2		status = 3	
	HR	95% CI	HR	95% CI	HR	95% CI
School*						
Mixed	0.441	0.246–0.794				
Classical	0.928	0.676–1.273				
Professional	0.300	0.173–0.521				
Outside Italy	1.486	0.407–5.427				
Technical	0.649	0.514–0.820				
Degree Course**						
L025					0.453	0.225–0.912
L026					1.370	0.838–2.239
Sex***						
Female			0.668	0.496–0.901		
School Grad.Mark	1.039	1.029–1.049	1.037	1.028–1.046	0.940	0.930–0.950
Time-by-Mark interaction			0.956	0.949–0.963	1.028	1.023–1.034

Reference categories: *Scientific, **L021, ***Male

a unit increase in graduation mark and a science-based school background both increase the probability of successful graduation (status = 1), while attending a technical, professional or mixed school tends to increase the probability of failing to complete the course.

For status 2 the school and the degree course attended were no longer significant, but gender was included in the model besides graduation mark and its interaction with time. The negative coefficient for being female reduced the subdistribution hazard, and the hazard ratio equal to 0.668 implies that, at any given time, females have a 33% lower cumulative incidence of the event compared to males, assuming other variables are constant. Moreover the effect of graduation mark on the subdistribution hazard changes over time, and is modelled as: $\beta(t) = 0.03637 - 0.04507 \times \log(\text{time})$.

A possible interpretation is that high-performing students are initially at greater risk (e.g., changing study programmes), but over time they stabilise or succeed, reducing risk.

For status 3 a significant predictor, besides graduation mark and its interaction with time, was the degree course attended. Students in group L025 had a 55% lower cumulative incidence of this event compared to those in the L021 category (HR = 0.453), suggesting a strong significant difference ($p = 0.027$) in risk between the L025 and L021 courses. In addition, the negative coefficient (-0.06218) of the graduation mark suggests that better school performance is associated with a reduced risk: for each unit increase in graduation mark, the subdistribution hazard decreases by about 6% (HR = 0.940) and this result is

highly significant ($p < 0.0001$). Moreover, the positive interaction term for time-varying effect (0.02783) indicates that the effect of graduation mark increases slightly over time, that is for each unit increase in graduation mark, the hazard ratio (HR = 1.028) increases by 2.8% over time. This suggests that the protective effect of graduation mark becomes somewhat stronger over time.

4 Conclusion

Failure to complete a course of university study is an important problem worldwide, representing a major waste of resources. It can be an issue even in Schools with highly competitive entry [5]. In this study we began an investigation of the problem of students' dropping out from three different degree courses in Economics of the University of Perugia, having a regional base (85.8%) and a success rate only just over 50%. A first analysis split the students into two groups: graduates and non-graduates. It identified graduation mark and type of secondary school attended as factors predictive of a positive outcome. Then, more deeply, as in other studies [16], we considered the different causes of dropping out, both applying high performance procedures which allow the identification of the most important factors, and implementing classical quantitative methods such as multinomial logistic regression and competing risk models. Even though the analyses lead to conceptually similar results, identifying the same important factors, various different facets have been distinguished.

In multinomial logistic regression (Table 5) we used as reference category the students who graduated, highlighting differences among various types of dropout. Conversely, in survival analysis one of the dropout causes (status = 0) naturally represents censored data. Therefore, the competing risk models (Table 6) focused attention separately on the probabilities of events 1, 2 or 3 considering the other two events in turn as competing. Results are therefore not directly comparable but of certain interest from an interpretative point of view. On the basis of the results obtained, the type of secondary school attended and graduation mark achieved seem to be the two factors that mainly influence the probability of academic success or dropout. Furthermore, the methodological differences between logistic regression and survival analysis allow us to detect how the passage of time influences the effect of some significant predictors in different directions for different causes of exit. The competing risk analysis, in addition, includes the gender variable, suggesting that females are less prone to change course (status = 2) compared to males.

However, this study will be improved in the near future by using data on annual student credits. It will then be possible to identify if there are any further points of attention, apart from those relating to the basic preparation acquired by students attending different types of secondary schools. The need to delve deeper into the topic will also lead to consideration of the cohorts of enrolment years after 2015.

References

1. Aina, C., Baici, E., Casalone, G., Pastore, F.: The determinants of university dropout: a review of the socio-economic literature. Socio-Econ. Plann. Sci. **79**, 101102 (2022). https://www.sciencedirect.com/science/article/pii/S003801212100094X
2. Anderson, J.J.: A vitality-based stochastic model for organism survival. In: Individual-Based Models and Approaches in Ecology, pp. 256–277. Chapman and Hall/CRC (2018)
3. Apumayta, R.Q., Cayllahua, J.C., Pari, A.C., Choque, V.I., Valverde, J.C.C., Ataypoma, D.H.: University dropout: a systematic review of the main determinant factors (2020–2024). F1000Research **13**, 942 (2024)
4. Boateng, E.Y., Abaye, D.A., et al.: A review of the logistic regression model with emphasis on medical research. J. Data Anal. Inf. Process. **7**(04), 190 (2019)
5. Caroni, C.: Graduation and attrition of engineering students in Greece. Eur. J. Eng. Educ. **36**(1), 63–74 (2011). https://doi.org/10.1080/03043797.2010.539676
6. Caroni, C.: Regression models for lifetime data: an overview. Stats **5**(4), 1294–1304 (2022). https://doi.org/10.3390/stats5040078
7. Cho, S.K., Mattke, S., Gordon, H., Sheridan, M., Ennis, W.: Development of a model to predict healing of chronic wounds within 12 weeks. Adv. Wound Care **9**(9), 516–524 (2020)
8. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc. Ser. B (Methodol.) **34**(2), 187–202 (1972)
9. Fine, J.P., Gray, R.J.: A proportional hazards model for the subdistribution of a competing risk. J. Am. Stat. Assoc. **94**(446), 496–509 (1999)
10. Fox, H., Topp, S.M., Lindsay, D., Callander, E.: A cascade of interventions: a classification tree analysis of the determinants of primary cesareans in Australian public hospitals. Birth **48**(2), 209–220 (2021)
11. Gupta, D., Gupta, D.: Telecommunication case study. In: Applied Analytics through Case Studies Using SAS and R: Implementing Predictive Models and Machine Learning Techniques, pp. 161–220 (2018)
12. Heisey, D.M., Patterson, B.R.: A review of methods to estimate cause-specific mortality in presence of competing risks. J. Wildl. Manag. **70**(6), 1544–1555 (2006)
13. Hosmer, D.W., Lemeshow, S.: Goodness of fit tests for the multiple logistic regression model. Commun. Stat. Theory Methods **9**(10), 1043–1069 (1980)
14. Huschka, T., et al.: 119 investigating factors influencing diagnostic safety in emergency department patients using decision trees. Ann. Emerg. Med. **82**(4), S52 (2023)
15. Jaffe, D., et al.: Chi-square automatic interaction detection (CHAID) analysis: understanding and informing treatment choice of second-line metastatic urothelial cancer therapy. Value Health **21**, S213–S214 (2018)
16. Kehm, B.M., Larsen, M.R., Sommersel, H.B.: Student dropout from universities in Europe: a review of empirical literature. Hung. Educ. Res. J. **9**(2), 147–164 (2019)
17. Menard, S.W.: Logistic Regression: From Introductory to Advanced Concepts and Applications. Sage (2010)
18. Ohno-Machado, L.: Modeling medical prognosis: survival analysis techniques. J. Biomed. Inform. **34**(6), 428–439 (2001)

19. Oyebanjo, M.O., Coker, O.M., Osaiyuwu, O.H.: Predicting the body weight of indigenous goat breeds from morphological measurements using the classification and regression tree (CART) data mining algorithm. Biotechnol. Anim. Husbandry **39**(1), 33–49 (2023)
20. Pho, K.H., McAleer, M.: Specification and estimation of a logistic function, with applications in the sciences and social sciences. Adv. Decis. Sci. **2**, 1–30 (2021)
21. Pierri, F.: Variable selection in binary logistic regression for modelling bankruptcy risk. In: Kitsos, C.P., Oliveira, T.A., Pierri, F., Restaino, M. (eds.) Statistical Modelling and Risk Analysis, pp. 143–155. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-39864-3_12
22. Pierri, F., Caroni, C.: Bankruptcy prediction by survival models based on current and lagged values of time-varying financial data. Commun. Stat. Case Stud. Data Anal. Appl. **3**(3–4), 62–70 (2017)
23. Putter, H., Schumacher, M., van Houwelingen, H.C.: On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: the Fine-Gray model revisited. Biom. J. **62**(3), 790–807 (2020)
24. Rasheed, M.: Analyzing applications and properties of the exponential continuous distribution in reliability and survival analysis. J. Positive Sci. **4**(5), 71–79 (2023)
25. Rodriguez, R.N.: Statistical model building for large, complex data: five new directions in SAS/STAT ® software. In: Proceedings of the SAS Global Forum Conference, Las Vegas, NV, USA, pp. 18–21 (2016)
26. SAS: SAS/STAT® 9.2 user's guide. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#phreg_toc.htm. Accessed 16 Mar 2025
27. SAS: SAS/STAT® 9.22 user's guide. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect004.htm. Accessed 16 Mar 2025
28. SAS: SAS/STAT® 9.22 user's guide. https://support.sas.com/documentation/onlinedoc/stat/141/hpsplit.pdf. Accessed 16 Mar 2025
29. Therneau, T., Crowson, C., Atkinson, E.: Using time dependent covariates and time dependent coefficients in the Cox model (2024). https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf
30. Venkatasubbu, P., Ganesh, M.: Used cars price prediction using supervised learning techniques. Int. J. Eng. Adv. Technol. (IJEAT) **9**(1S3), 216–223 (2019)
31. Wang, P., Zheng, X., Li, J., Zhu, B.: Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos, Solitons Fractals **139**, 110058 (2020)
32. Zelenkov, Y.: Bankruptcy prediction using survival analysis technique. In: 2020 IEEE 22nd Conference on business informatics (CBI), vol. 2, pp. 141–149. IEEE (2020)