doi.org/10.26398/asaproc.0120

Early detection of university dropouts: An experiment using video-lectures access data

Maria Cristiana Martini
Department of Communication and Economics, University of Modena and Reggio Emilia, Reggio Emilia, Italy.

1. Introduction

University attrition is one of the most complex and adverse events in students' career. The percentage of students who do not complete a bachelor's degree program within its theoretical duration ranges from less than 40% in Israel and the United Kingdom to 80% or more in Colombia, the French Community of Belgium and Italy (OECD, 2022). If we enlarge the time interval to three years after the theoretical duration, the non-completion rate still ranges from 15% in United Kingdom to 50% in Brazil. In Italy 47% of students who entered a bachelor's programme did not graduate three years after the theoretical duration, and 32% are no more enrolled in tertiary education (OECD, 2022).

Attrition is a potentially disastrous event in a student's life, and it also impacts negatively the university. Early identification of students who are prone to abandon their studies allows specialized professionals to implement action plans for retention. Machine learning techniques have widely been used to predict university dropouts (see, e.g., Agrusti et al., 2019 for a review). Student dropout predictions are usually made on the basis of features that depend both on the determinants of university dropouts (see Aina et al., 2022 for a review) and the availability and timeliness of each information. Most predictions are then based on administrative data, like personal information, family background and study career (school background and academic performance), that are costless and timelier, while ad-hoc surveys on involvement and intentions can be more precise but are expensive and difficult to obtain. As students advance through their curriculum, more data are available, but dropouts are less frequent, and interventions become less effective.

The availability of online courses can provide a further source of possible features to predict dropouts, since the video-lectures are often accessed through individual login credentials and these operations produce large log datasets, giving information on time and frequency of video-lectures fruition.

The aim of this paper is to integrate data on video-lectures usage during the first year at university in the prediction of students' dropout, since these data are automatically collected, easily available, and offer an insight on study methods and dedication. Section 2 describes the data and the methodology; Section 3 presents some results, while conclusions are drawn in Section 4.

2. Data and methods

Classes at the Department of Communication and Economics of the University of Modena and Reggio Emilia have been recorded and made available to students starting from September 2017, when the ONELab project was launched (Furini et al., 2020); face-to-face classes were regularly held (with the exception of the pandemic period), but most students accessed the video-lectures to supplement their preparation and review some difficult passage.

The log data is a large dataset where each record corresponds to an access to the ONELab platform, and reports information on the day and time of the access and the accessed course; these data have been aggregated to gather information on the time, type and frequency of accesses for each student.

The data for this study integrates two data sources: log data on video-lectures fruition in the first

months of the first year (from September to the end of May) and administrative data on students' demographics and study career. Data include all the freshmen of the five degree courses of the Department in the academic years 2017/18 and 2018/19, except those who withdrew before June in the enrolment academic year. This exclusion has a double reason: first, the importance of data on video-lectures fruition and academic achievement on dropouts would be magnified, since early dropouts do not pass exams and do not watch video-lectures; furthermore, early dropouts are not interesting to predict, since they are officially known before the data to forecast them is available.

After the deletion of early dropouts, the dataset consisted of 2,269 observations (1,058 enrolled in 2017, 1,222 in 2018), but five were discarded: one died few months after the enrolment, and two couples of homonymous students were enrolled to the same study program, making it difficult to correctly merge their administrative data to the log data. The reason for choosing students enrolled in 2017 and in 2018 is twofold: avoiding the pandemic period, when lectures fruition was completely upset, and including students who have had plenty of time to graduate, so that dropouts' definition becomes simpler.

The first encountered problem, in fact, is the difficulty to define dropouts in an unambiguous and consistent way. Administrative data recorded 395 official withdraws occurring after May in the first University year out of 2264 observations (17.4%), but these were only a part of all the actual dropouts, since many students simply stopped attending classes and paying taxes, but do not withdraw at the administrative office. However, students who stop paying taxes do not always express a final will to dropout: some students might pay late, or might interrupt their studies for a period, due to economic or family problems. After analysing the enrolment status in each academic year up to 2022/23, students have been defined as dropouts if they did not graduate, and they did not enrol to university for at least the last two years: these are 649 cases (28.7%), while the non-completion rate (dropouts and still enrolled students) is 44.3%, and both data are consistent with the OECD (2022) data for Italy.

The final dataset included pseudo anonymized data resulting from the merge of the administrative data and the aggregated log data; besides the outcome variable on dropouts, this dataset included a large number of students' characteristics which can be useful features in a machine learning prediction:

- Basic features, i.e. personal information, family background and school background: gender; age; nationality; specific learning disorders; place of residence; tax area; high school type; high school grade; selected degree program. These features are already available when a student enrolls at the university and can be immediately used to predict dropouts.
- Academic performance in the first semester: number of ECTS, number of passed exams and average grade after the first semester. These features are available after the end of February, when the winter exams session is over.
- Video-lectures usage[1]: total number of accesses; number of courses accessed at least 1, 5, or 10 times; month of first access; number and percentage of accesses per time slot; number of accesses to the favorite course and in the favorite time slot. The availability of these features depends on the time period one wants to consider for video-lectures fruition: this analysis takes into account video-lectures usage between September and May, so these features are available from the month of June. Considering a shorter period implies that data will be available earlier, but will likely be less informative, while a longer observation period implies more informativeness but a delayed data availability.

Some data pre-processing operations have been carried out before data analysis. The dataset has been split in an 80% training data, and a 20% testing data. Although there are very few missing

[1] Since early dropouts are excluded from the dataset and only dropouts occurring after May are considered, the video-lectures usage is limited to the accesses before June.

data, these have been replaced by means of k-nearest neighbours (K-nn) technique. All the features have then been standardized, and One-Hot encoding procedure has been applied to categorical variables. The response variable is dichotomous (dropout/non dropout); hence the prediction regards a binary classification problem. Alternative machine learning algorithms and models have been applied, among the most commonly used in analogous classification problems:

- Decision trees according to the original Classification and Regression Trees (CART) algorithm that uses binary splits (see Breiman et al., 1984);
- Random forest (RF), that is an improved version of CART generated by bootstrapping on the training sample and averaging over the multiple versions of the tree (see e.g. Breiman, 2001);
- K-nearest neighbours (KNN), a basic non-parametric supervised learning method (see Cover and Hart, 1967);
- Neural networks (NN), a machine learning model inspired by the functioning of the human brain, that involves one or more levels of hidden layers; here a simple feed-forward neural network with a single hidden layer is applied (see, e.g., Ripley, 1996);
- Naïve Bayes (NB), a simple linear probabilistic classification algorithm that utilizes the Bayes rule and assumes conditional independence of the features (for a discussion on its optimality under wide violations of the assumption see, e.g, Domingos and Pazzani, 1997).

Each model was estimated on the training dataset, and the hyperparameters were tuned based on a maximization of the area under the ROC curve. Then, each model is assessed on the testing data, with a 10-fold repeated cross-validation. Pre-processing and analyses have been carried out with the *caret* package of R (Kuhn, 2008).

3. Results

Accuracy, sensitivity and specificity are calculated as averages of the measures obtained for each attempt and are reported in Table 1.

Table 1. Accuracy, sensitivity and specificity for the alternative classification methods using different feature sets.

Feature set	Performance measure	CART	RF	KNN	NB	NN
Basic features	*Accuracy*	0.692	0.724	0.677	0.659	0.728
	Sensitivity	0.374	0.503	0.303	0.006	0.503
	Specificity	0.859	0.838	0.872	1.000	0.845
Basic features and academic achievements	*Accuracy*	0.768	0.774	0.721	0.675	0.792
	Sensitivity	0.523	0.574	0.419	0.052	0.600
	Specificity	0.896	0.878	0.879	1.000	0.892
All features	*Accuracy*	0.770	0.790	0.730	0.704	0.777
	Sensitivity	0.548	0.561	0.432	0.600	0.580
	Specificity	0.886	0.909	0.886	0.758	0.879
ONELab usage only	*Accuracy*	0.693	0.695	0.675	0.664	0.684
	Sensitivity	0.445	0.361	0.394	0.465	0.400
	Specificity	0.822	0.869	0.822	0.768	0.832

Neural networks perform slightly better than the other methods with most of the feature sets, but in general the results are quite disappointing, especially regarding the most important outcome, i.e. sensitivity, the power to correctly detect dropouts. Sensitivity, in fact, is almost null for the naïve Bayes algorithms that do not use video-lectures usage features, but also with the other methods it ranges from 0.303 (k-nearest neighbors, basic features) to 0.6 (for neural networks on basic features

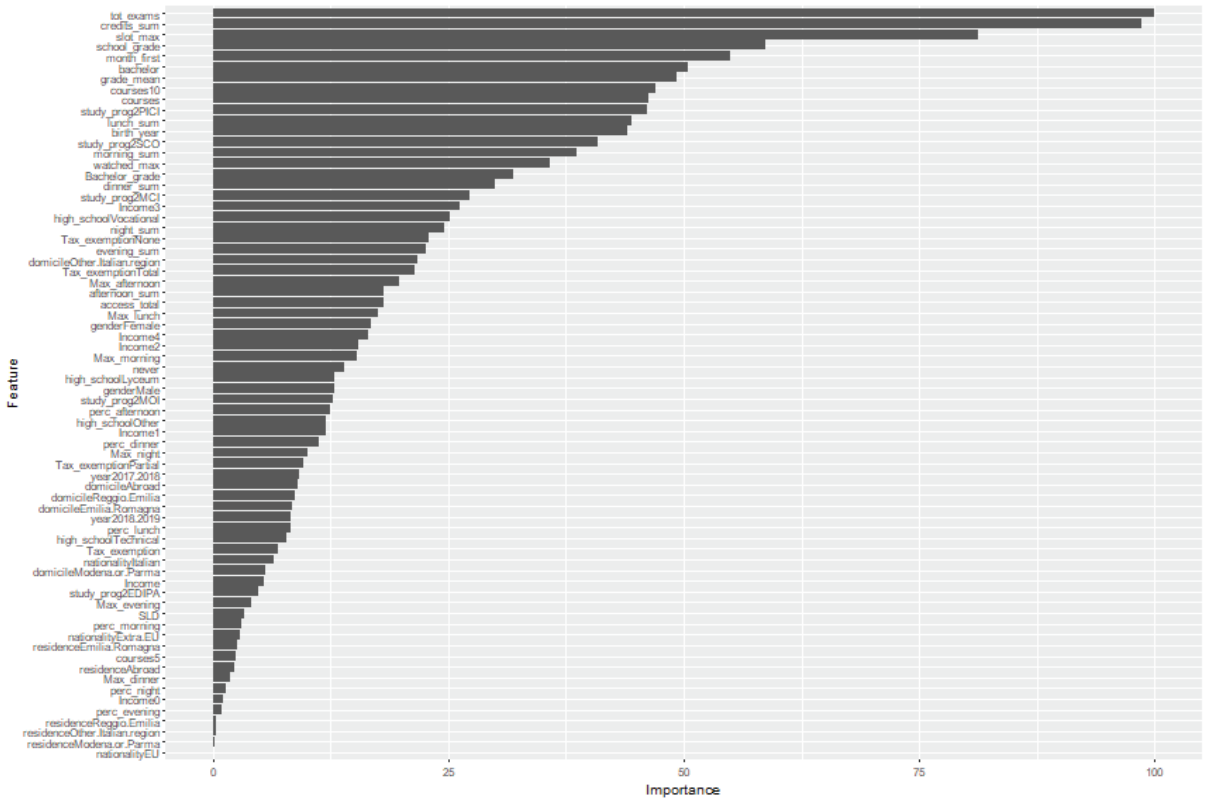
and academic achievements, and for naïve Bayes with all features). Moreover, accuracy is never high, while the somehow higher level of specificity can probably be explained by the slightly imbalanced structure of the data. Dropouts, in fact, are less than 30% of the observations, and consequently the trivial decision to always predict "no dropout" would guarantee a specificity larger than 0.7.

The comparison between performance measures with basic features and academic achievements, and performance measures with all features indicates that ONELab access data add very few (or nothing) to the usual features from administrative data. On the other side, in many cases the use of video-lectures usage only leads to the same performance as the use of the much larger and complex set of administrative data indicated as "basic features".

The most accurate and sensitive predictions (with the exception of the naïve Bayes) are those which include data on the academic achievements during the winter session of exams, even though this group of features only includes three variables, while the use of the large groups of basic features and ONELab usage features alone does not lead to outstanding results.

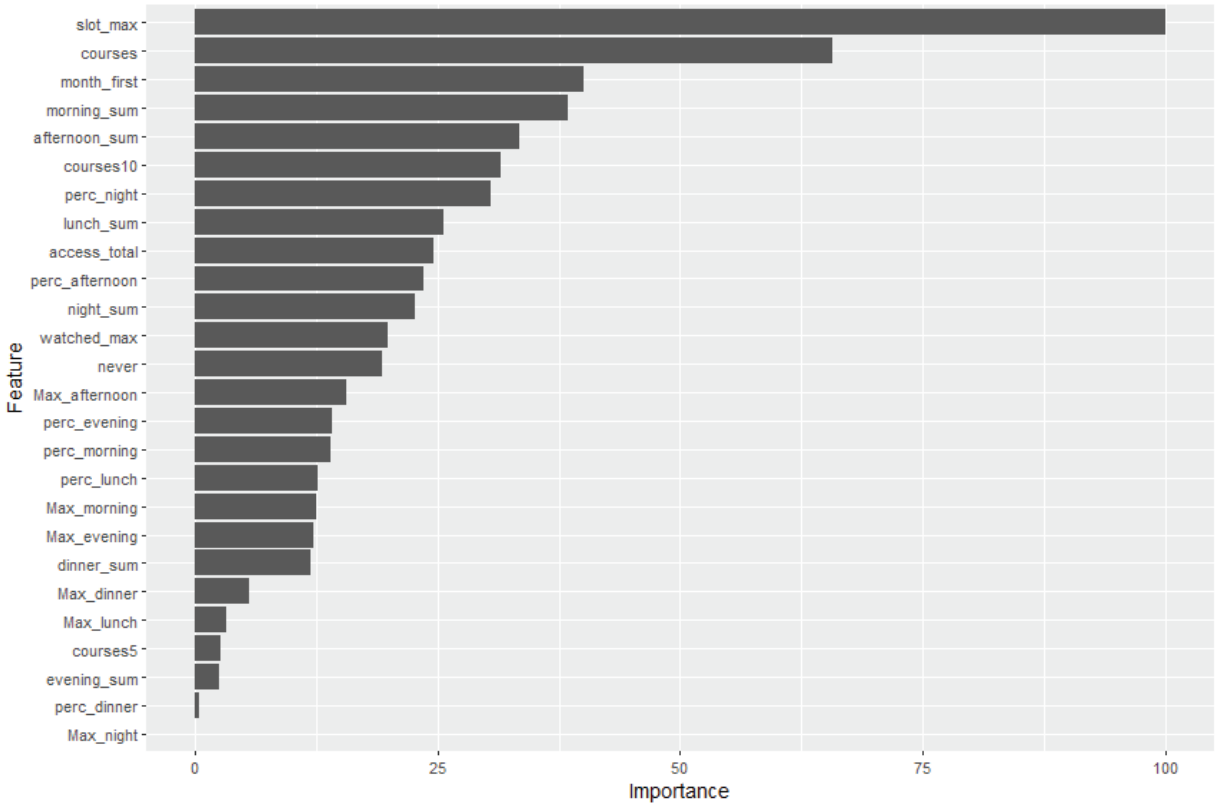
The importance of the academic achievement variables is confirmed if we look at the variable importance of the whole set of features with the neural networks method in Figure 1 (the variable importance obtained with the other algorithms is analogous); here we notice that the first positions are occupied by academic achievement variables ('tot_exams' and 'credits_sum'), followed by the number of video-lectures accessed in the favourite time slot ('slot_max'). The third feature on academic achievements is also among the ten most important variables.

Figure 1. Variable importance with neural network method; all features.



Limiting the variable importance analysis to the video-lectures usage features (Figure 2), the most important features are the number of lectures accessed in the favourite time slot ('slot_max'), and the number of different courses that have been accessed at least once ('courses'), which can probably capture the extent of time devoted to study. Some of these features, as well as the complete set discussed above, are partially redundant, and this impacts the differences in variable importance.

Figure 2. Variable importance with neural network method; video-lectures usage.



4. Conclusions

The general behaviour of the predictive models was in a sense disappointing, not especially with regard to the predictive power of the proposed variables on online access to lectures recordings, but for the overall scarce capacity to predict dropouts, even with traditional administrative and career data.

If compared to the results of similar studies (Del Bonifro et al., 2020; Berens et al., 2018) accuracy and sensibility are a bit lower. What is different in this study? First, in order to think of an early prediction, this study only considers the academic achievements obtained during the first exam session (January and February), while other studies include all the credits obtained in the first year and sometimes (for the prediction of dropouts occurring after the first year) also in the next years. In fact, although these analyses have not been shown here, data predictions based on the achievements of the whole academic year are a bit better but, at the time of the last exam session, a large amount of dropouts have already happened, arising the problem of a mutual dependency between dropouts and some features. On the other hand, when considering only enrollment or first semester data, also Berens et al. (2018) obtain accuracy levels lower than 0.7 and (in some cases) sensibilities under 0.5.

Moreover, in the attempt to reproduce the conditions in which a prediction predicts, in this study all the students who withdrew from university before June were excluded from the analyses. It's only 78 cases out of the original 2,342, but it would have raised the percentage of dropouts over 35% and, above all, these early dropouts were likely to be easier to detect (almost no exams, no credits, no access to the video-lectures), and they would have been about 10% of the total dropouts. But, again, we cannot call "prediction" the announcement of something that has already happened at the prediction time.

When adding the features on video-lectures fruition, the improvement was mostly negligible, and this is bad news. However, these data are costless, because they are automatically collected when students log in to access the lectures recordings, and harmless. Furthermore, the integration

of these features was useful to enhance the sensitivity in those classification models where sensitivity was alarmingly low. Even more promising: in most cases the predictions based only on ONELab access data are not worse than the predictions based on the whole set of features (except the academic achievement).

The early prediction issue remains at stake: is it actually possible to make an effective early prediction? Maybe a very early prediction, even before the winter session of exams, when students still have time to recover and learn how to study, or re-orientate to a different study programme? According to these results, and looking at the comparison to the results of other studies, when it comes to dropout predictions there is a trade-off between the need to detect students at risk as early as possible, and the growing inaccuracy of detection. Maybe, in the future perspective of a "University 4.0", students will be traced more steadily (monitoring their presence in the classroom, in the studying rooms, their use of online textbooks and exercises, and so on), and very early detection will come true. Or, who knows, it's mere science-fiction.

References

Agrusti, F., Bonavolontà, G., Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-Learning and Knowledge Society*, **15**(3), pp. 161-182.

Aina, C., Baici, E., Casalone, G., Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, **79**, pp. 1-16.

Berens, J., Schneider, K., Görtz, S. Oster, S., Burghoff, J. (2018), Early detection of students at risk – Predicting student dropouts using administrative student data and machine learning methods. CESifo Working Paper No. 7259, CESifo, Munich.

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

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, (CA).

Cover, T.M., Hart, P.E. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**(1), pp. 21-27.

Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P. (2020). Student dropout prediction, in I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, E. Millán (eds). *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science*, vol 12163. Springer, Cham.

Domingos, P., Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, **29**, pp. 103-130.

Furini, M., Galli, G., Martini, M.C. (2020). An online education system to produce and distribute video lectures. *Mobile Networks and Applications*, **25**, pp. 969-976.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, **28**(5), pp. 1–26.

OECD (2022). *Education at a Glance 2022: OECD Indicators*. OECD Publishing, Paris.

Ripley, B.D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.