Mitigating academic dropout through ML: The case study of the University of Florence

Stefano Masini[a], Silvia Bacci[b], Alessandra Petrucci[b], Bruno Bertaccini[b]
[a] Department of Computer Science, University of Pisa, Pisa, Italy.
[b] Department Statistics, Computer Science, Applications, G. Parenti University of Florence, Florence, Italy.

1. Introduction

In Italy, student dropout represents one of the most critical challenges facing the higher education system. According to the ANVUR 2023 Report (AA.VV., 2023), the dropout rate between the first and second year of bachelor's degree programs has shown a significant increase for students enrolled in the academic year 2020/21, rising to 14.5%, thereby returning to levels observed a decade ago. Dropout risks, however, are not limited to the first year: 20.4% of students abandon their studies within three years of enrolment in bachelor's programs, increasing to 24.2% after six years. At the University of Florence, first-year dropout rates for bachelor's programs exceed the national average, although a slight decline has been observed over the last four years (from 27.1% in 2020/21 to 25.3% in 2023/24). Contributing factors include insufficient academic preparation, lack of motivation, economic difficulties, and inadequate guidance, all resulting in a misalignment between student expectations and the academic experience.

Dropout entails not only a substantial cost in terms of public and private resources, but also negatively affects human capital development and delays entry into the labor market. Given this context, the adoption of effective preventive policies, such as targeted pre-admission guidance and ongoing academic tutoring, is increasingly urgent.

In this paper we develop a predictive system capable of estimating the probability that students will drop out by the end of their first academic year. The proposed system relies on an ensemble strategy that, accounting for various individual features, combines two machine learning (ML) models, specifically, XGBoost and TabNet.

The rest of the contribution is organized as follows: in Section 2, we describe the data; in Section 3, we illustrate the methodological approach to develop the predictive system; in Section 4, we provide evidence of the main results. Section 5 concludes with some final remarks.

2. Data description

Data used to train and test the ML system are collected from the administrative archives of the University of Florence. The dataset includes 59,165 students enrolled during the academic years 2020/21 through 2023/24 in one of the following degree programs: bachelor's degrees (L), master's degrees (LM), or single-cycle master's degrees (LMCU). These programs are offered across the ten schools of Florence University: Agronomy, Architecture, Economics, Law, Engineering, Psychology, Health sciences, Mathematical, physical and natural sciences, Political sciences, and Humanities and education.

All the degree programs suffer some dropout, although with significant differences among schools and types of degree programs. In the following, we distinguish between *strong* dropout, when a student drops out of the University of Florence, and *weak* dropout, when a student transfers to another degree program (either within or outside the University of Florence). Table 1 shows the rates of strong dropout registered by the end of the first academic year.

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Stefano Masini, Silvia Bacci, Alessandra Petrucci, Bruno Bertaccini, *Mitigating academic dropout through ML: The case study of the University of Florence*, pp. 297-302, © 2026 Author(s), CC BY 4.0, DOI 10.36253/979-12-215-1062-1.52

Table 1. Relative frequencies of strong dropouts by the end of the first academic year, per type of degree program, cohort of enrolment, and academic school.

	L				LM				LMCU			
	20/21	**21/22**	**22/23**	**23/24**	**20/21**	**21/22**	**22/23**	**23/24**	**20/21**	**21/22**	**22/23**	**23/24**
Agronomy	0.39	0.41	0.39	0.35	0.16	0.16	0.09	0.13				
Architecture	0.27	0.22	0.27	0.20	0.14	0.15	0.11	0.12	0.24	0.18	0.26	0,.4
Economics	0.27	0.30	0.30	0.27	0.15	0.12	0.12	0.09				
Law	0.08	0.07	0.07	0.06				0.35	0.19	0.25	0.25	0.22
Engineering	0.28	0.35	0.29	0.29	0.06	0.09	0.04	0.06				
Psychology	0.14	0.14	0.12	0.12	0.07	0.08	0.09	0.05				
Health	0.18	0.19	0.20	0.20	0.10	0.14	0.13	0.09	0.12	0.12	0.11	0.1
Math.Phys.Nat.	0.38	0.41	0.37	0.33	0.08	0.05	0.08	0.07				
Political sc.	0.34	0.32	0.29	0.27	0.14	0.17	0.18	0.17				
Human.Educ.	0.30	0.30	0.28	0.30	0.13	0.15	0.13	0.12	0.07	0.11	0.06	0.12

The highest dropout rates are observed for the bachelor's degrees, in particular for the degree programs of the schools of Mathematical, physical and natural sciences (0.33-0.41), Agronomy (0.35-0.41), Engineering (0.28-0.35), Humanities and education (0.28-0.30), Political sciences (0.27-0.32), and Economics (0.27-0.30). Notable exceptions are represented by the bachelor's degree programs with limited enrolment, that is, those offered by the schools of Psychology (0.12-0.14) and Health (0.18-0.20). As expected, lower dropout rates are observed for master's and single-cycle master's degrees. Clearly, weak dropout rates (not reported here) are slightly higher than strong ones because those who drop out of a degree program may not drop out of the university, but the reverse is not true.

For each student, administrative data on their academic career are available, including the degree program attended (with details such as type and academic school), year and age at enrolment, the outcome of the entry knowledge assessment test (if applicable), and the exams taken by the end of the first year, along with their course typology (A: foundational, B: characterizing, C: related, D: elective, F: other, I: non-graded pass/fail), credits, dates, and grades.

Moreover, the administrative archives of the university provide information related to the students' characteristics. We use these characteristics as additional features in the ML system to improve its predictive capability. In detail, we enrich the dataset with type of high school attended by the student, high school final score, age of the student upon completion of high school, geographical area (Italian region) where the student has lived during high school, student's job condition, student's socio-economic status, and the student's right to benefit from compensatory measures due to certified disabilities.

3. Methodological approach

Student dropout prediction is a classic tabular ML problem with heterogeneous feature types, missingness, nonlinear interactions, and moderate class imbalance (few dropouts relative to non-dropouts). No single learner dominates across such conditions; different algorithms capture different statistical regularities and therefore make complementary errors. Combining them can reduce variance and bias simultaneously and improve generalization, especially when base learners are diverse (heterogeneous inductive biases). This principle is well established in ensemble learning research (Wolpert, 1992; Caruana et al., 2004).

In our setting, we empirically compared gradient-boosting decision trees like XGBoost (Chen and Guestrin, 2016), LightGBM (Ke G. et al., 2017), and CatBoost (Prokhorenkova et al., 2018), a multilayer perceptron (Almeida, 2020), and a deep architecture (TabNet) tailored to tabular data (Arik and Pfister, 2021). We selected XGBoost and TabNet as base learners for the deployed ensemble because they provided strong complementary performance: XGBoost is a scalable, regularized gradient-boosted decision tree framework, with excellent handling of heterogeneous tabular features, and it naturally supports instance weighting – useful for addressing class imbalance - via class weigths; TabNet is a deep architecture designed specifically for tabular data, which employs a

sequential mechanism to perform sparse feature selection and capture complex conditional interactions across heterogeneous inputs – patterns that tree-based models and standard multi-layer perceptrons (MLPs) often fail to capture effectively. Empirically, the two models produced partially uncorrelated error patterns on our data, improving ensemble potential (Dietterich, 2000).

To combine the output of the two models, we adopt a blending scheme that linearly combines calibrated dropout probabilities from XGBoost and TabNet:

$$\text{Final probability} = \alpha\ XGBoost_prob + (1 - \alpha)\ TabNet_prob.$$

The final classification (i.e., dropout or not) is then determined by comparing this probability to a threshold k. If the probability exceeds k, the student is classified as a dropout; otherwise, as a non-dropout. This simple approach represents a strong and widely used baseline in ensemble learning – both in academic research and competitive settings – often matching or outperforming more complex approaches when the underlying base learners (here, XGBoost and TabNet) are well tuned and properly calibrated (Wolpert, 1992).

To determine the optimal values for the weight α and the threshold k, we define two search grids: α ranged from 0 to 1, and k ranged from 0.3 to 0.7. Predictions are computed for each pair (α, k), and the combination achieving the highest accuracy is selected for the final model.

To extract the most from each learner, we optimize base-model hyperparameters independently using cross-validation on the training data. Each model is trained with instance weights or class weights to partially offset class imbalance. This "optimize-then-blend" approach is consistent with ensemble selection practice: build strong, diverse members first, then combine (Caruana et al., 2004).

Finally, to ensure transparency and interpretability, we analyze the final ensemble with SHAP (SHapley Additive exPlanations). SHAP (Lundberg and Lee, 2017) attributes each prediction to input features using game-theoretic, where each feature is treated as a *player* contributing to the model prediction. By computing Shapley values, SHAP quantifies the contribution, which can be positive or negative, of each feature to a specific prediction (local explanation) and can summarize their overall importance across the dataset (global explanation). For tree ensembles like XGBoost, TreeSHAP yields exact, fast Shapley values; for TabNet, kernel or sampling-based SHAP can be used, and TabNet's inherent feature masks offer an additional interpretable signal about which features were attended to at each decision step. This dual perspective makes it possible to identify both the most influential factors and the direction of their effect, enhancing trust in the system's outputs.

4. Experimental setup and results

We train the predictive dropout system following two parallel experiments based on a different definition of the training set. In the first experiment, we consider the entire dataset of the four academic years (2020/21, 2021/22, 2022/23, 2023/24) and then, we split it into a training set and a test set, randomly assigning observations to one or the other in a ratio of 80 (training set) to 20 (test set). In the second experiment, we train the model on the first three academic years (2020/21, 2021/22, 2022/23) and test it on the last year (2023/24). Under the first experiment, the size of the resulting test set is 11,833 students (belonging to the four academic years), whereas in the second experiment, the test set size equals 14,502 students enrolled on the 2023/24 academic year.

In both experiments, the ML system achieved very high accuracy: 91.9% in the first experiment and 91.3% in the second for strong dropout prediction, and slightly higher values (92.9%–92.7%) for weak dropout prediction. Table 2 reports the values of accuracy detailed per academic school and type of degree program (L, LM, LMCU), obtained with the second experiment.

Table 2. Accuracy per academic school and type of degree program (L, LM, LMCU): Test set size, accuracy for prediction of strong dropout, accuracy for prediction of weak dropout (second experiment).

	L			LM			LMCU		
	Test set size	Strong dropout	Weak dropout	Test set size	Strong dropout	Weak dropout	Test set size	Strong dropout	Weak dropout
Agronomy	314	0.929	0.921	118	0.942	0.910			
Architecture	500	0.930	0.919	169	0.938	0.924	268	0.939	0.924
Economics	1615	0.925	0.900	377	0.930	0.922			
Law	877	0.898	0.883	20	0.982	0.982	516	0.921	0.921
Engineering	1198	0.937	0.921	390	0.900	0.891			
Psychology	467	0.934	0.924	285	0.937	0.918			
Health	1034	0.928	0.912	538	0.931	0.926	556	0.924	0.920
Math.Phys.Nat.	931	0.929	0.911	288	0.924	0.923			
Political sc.	721	0.924	0.919	246	0.926	0.917			
Human.Educ.	2245	0.923	0.915	625	0.922	0.914	204	0.921	0.909

To evaluate how much the academic features affected the system's predictive ability, Figure 1 shows the SHAP values calculated for each feature.

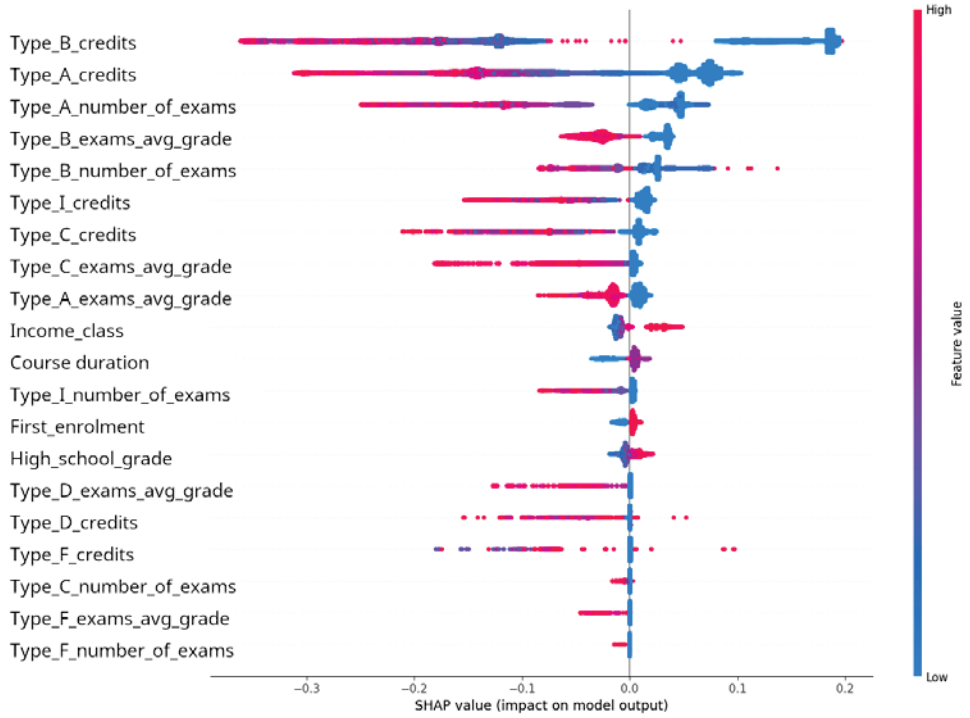

Figure 1. SHAP summary: Impact of individual features on the dropout probability.

In the SHAP summary plot, each dot represents a student. The x-axis shows the SHAP values computed for each student. SHAP values are expressed in the same units as the model outputs. Since our ensemble returns probabilities, the SHAP values represent the difference between the predicted probability and the baseline probability. The color (feature value) immediately shows the direction of the effect. For example, if most red points lie to the left of zero for a given feature, this means that high feature values reduce the predicted dropout probability. Conversely, if most red points lie to the right of zero, high values will increase the dropout probability. Furthermore, a wide spread (i.e. a large horizontal range) indicates that the feature has a variable effect across instances (i.e. strong discriminative power), whereas a tight, thin distribution near zero suggests that the feature has limited influence on the prediction.

On the y-axis, features are sorted by mean absolute SHAP, so the top rows have the largest average impact on predictions.

Thus, SHAP values provide both global and local interpretability. At the global level, feature importance highlights the most influential predictors, confirming that indicators of academic engagement (e.g., number of exams, average exam grades, and credits earned), together with program characteristics, constitute critical risk factors. The ranking of features along the y-axis in the summary plot directly reflects this global perspective, as it addresses the question: *which features contribute most to the model's predictions on average across the entire population?*

At the local level, SHAP values enable a case-by-case interpretation of predictions, thereby supporting personalized student guidance and targeted interventions. Each point in the plot corresponds to an individual student and represents the contribution of a specific feature to that student's prediction. In this way, SHAP clarifies *why* the model assigned a particular score, and analyzing the distribution of local contributions facilitates a detailed understanding of individual risk profiles.

In our analysis, the most influential features include *Type_B_credits, Type_A_credits, Type_A_number_of_exams, Type_B_exams_avg_grade,* and *Type_B_number_of_exams*, among others. These variables exhibit the largest average impact on the model's output, as reflected by their mean absolute SHAP values. The horizontal range of SHAP values (approximately −0.35 to +0.2) indicates that their effect on predicted dropout probability is non-negligible. For instance, a SHAP value of −0.3 represents a substantial negative contribution, decreasing the predicted probability of dropout by roughly 30 percentage points.

Finally, the color gradient encodes feature values and reveals the direction of their influence. For example, in the case of *Type_B_credits*, the concentration of red points on the left side of the axis suggests that higher values of this feature tend to reduce the model output - meaning that the more credits a student has earned, the lower the predicted probability of dropping out.

5. Conclusions

This study proposes a decision-support tool for university governance, aimed at the early identification of students at risk of dropping out. A machine learning model has been developed based on an ensemble of XGBoost, well-suited for handling complex problems with strong variable interactions, and TabNet, which provides advanced neural capabilities and selective attention mechanisms.

Trained on enrolment cohorts at the University of Florence from the 2020/21 to the 2022/23 academic years, the model has shown an exceptionally level of accuracy (approximately 92%) in predicting outcomes observed in the 2023/24 academic year across all degree programs. This result is particularly remarkable given that the earliest training cohort coincided with the COVID-19 pandemic, a period marked by higher dropout rates compared to subsequent cohorts.

Future research will pursue two main directions aimed at further enhancing the predictive capacity of the proposed system. We will explore adaptive strategies to maximize predictive performance, with a particular emphasis on sensitivity (recall) for the dropout class. Prioritizing recall is crucial in this application domain, as failing to identify students at risk may have more severe consequences than generating false positives.

Moreover, we plan to enrich the ensemble architecture by integrating additional machine learning algorithms beyond the current selection, thereby capturing complementary data representations and reducing model variance. While the present study relies on a simple weighted averaging scheme to combine base learners, future work will evaluate the effectiveness of meta-models (e.g., logistic regression or neural networks) to learn the optimal combination of base predictions in a data-driven manner. Such an approach is expected to improve particularly the recall of dropout predictions, while maintaining an appropriate trade-off with precision and overall accuracy.

References

AA.VV. (2023). *Rapporto sul sistema della formazione superiore e della ricerca 2023*. ANVUR. https://www.anvur.it/sites/default/files/2024-12/Sintesi-Rapporto-ANVUR-2023.pdf

Almeida, L.B. (2020). Multilayer perceptrons. In E. Fiesler, R. Beale (eds.). *Handbook of Neural Computation*. CRC Press. pp. C1-C2.

Arik, S., Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35, pp. 6679-6687.

Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML '04)*. Association for Computing Machinery, New York, NY, USA, 18. Doi: 10.1145/1015330.1015432

Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. Doi: 10.1145/2939672.2939785

Dietterich, T.G. (2000). Ensemble methods in machine learning. In: *Multiple Classifier Systems*. *MCS 2000. Lecture Notes in Computer Science*, vol. 1857. Springer, Berlin, Heidelberg. Doi: 10.1007/3-540-45014-9_1

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook (NY, USA): Curran Associates Inc., pp. 3149–3157.

Lundberg, S.M., Lee, S. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook (NY, USA): Curran Associates Inc., pp. 4768–4777.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Red Hook (NY, USA): Curran Associates Inc., pp. 6639–6649.

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks,* 5(2), pp. 241-259.