MOX-Report No. 57/2025

A Novel DNA-Inspired Framework to Study University Dropout: Insights from Politecnico di Milano

Guagliardi, O.; Masci, C,; Breschi, V; Paganoni A, ; Tanelli, M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it https://mox.polimi.it

A Novel DNA-Inspired Framework to Study University Dropout: Insights from Politecnico di Milano

Oriana Guagliardi[a], Chiara Masci[b], Valentina Breschi[c], Anna Maria Paganoni[d] and Mara Tanelli[a]

[a]Department of Electronics, Information and Bioengineering, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133, Milano, Italy; [b]Università degli Studi di Milano, Via Festa del Perdono 7, 20122 Milano, Italy; [c]Eindhoven University of Technology, 5600 MB Eindhoven, Netherlands; [d]Department of Mathematics, MOX– Modelling and Scientic Computing, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133, Milano, Italy

Abstract

This study presents Dropout-DNA, a novel data-driven tool designed to assess university dropout risk by profiling students through a combination of early indicators and academic progress. The approach emphasizes the need for context-aware and interpretable models in predicting student dropout, offering a significant advancement in the field of student retention analytics. Results show that while early indicators are valuable, incorporating academic performance significantly enhances predictive accuracy. The model, although generalizable across engineering courses, performs best when tailored to the specific degree program it was trained on. This finding underlines the importance of adapting predictive tools to the unique characteristics and dropout patterns of individual study programs. The practical implications are considerable: by identifying at-risk students early, institutions can implement targeted and personalized interventions, improving the effectiveness of student support services. The Dropout-DNA's quantifiable representation of risk allows for more strategic policy-making at the institutional level. Looking ahead, future research will focus on the temporal evolution of dropout risk profiles, enabling dynamic, time-sensitive monitoring and intervention throughout the academic journey.

Keywords

Student Profiling; Student Dropout Prediction; Statistical Modeling; Machine Learning Algorithms

---

Corresponding author(s):

Oriana Guagliardi Department of Electronics, Information and Bioengineering, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133, Milano, Italy. Email: oriana.guagliardi@polimi.it

Introduction

The dropout phenomenon is one of the most critical challenges in the university setting. Even if much progress was made concerning access to higher education and enrolment rates are now higher, the dropout rate remains a relevant issue. Eurostat's latest statistic (2022) reports a 10% rate in the EU. In particular, Italy faces a concerning challenge, having one of the lowest tertiary graduation rates among OECD countries. According to OECD data, Italy's tertiary graduation rate lags behind the OECD average, reflecting difficulties in retaining students and guiding them to completion.

This has wide-ranging consequences: for students, it often leads to lost educational and career opportunities, and psychological distress (Berka and Marek, 2021; Skrbinjek et al., 2024). For institutions, it can compromise funding and academic quality (Zotti, 2015). For society, it results in inefficient use of public resources and a shortage of skilled professionals (Ghignoni, 2017; Lorenzo-Quiles et al., 2023; Skrbinjek et al., 2024).

The need to understand this phenomenon to intervene and reduce its numbers has led to many policies and interventions by universities and governments. Focusing on Italy, the Ministry of University and Research (MUR) is implementing initiatives to improve academic guidance and student retention. One is a digital platform designed to facilitate students' understanding of the Italian academic system, providing comprehensive information on courses, accommodation and scholarships. To bolster the right to education and mitigate financial burdens, the Italian government has earmarked nearly 1 billion euros in the budget law: 500 million to extend PNRR scholarships until 2026 and over 400 million for university housing (Ministero dell'Università e della Ricerca, 2023). In addition, various solutions are adopted at the individual university level. The Politecnico di Milano, for instance, has incorporated specific policies into its 2023-2025 Sustainability Strategic Plan. These include support for students with disabilities and learning disorders, tutoring, welfare services such as cultural, food, travel and sustainable mobility agreements, and off-campus spaces for co-learning and co-creation.

While these policies are promising, dropout is multifaceted and demands a holistic approach. Numerous studies highlight the interplay of individual, academic, and systemic factors. The literature distinguishes between theoretical and data-driven approaches. Theoretical models, such as those by Spady and Tinto (Spady, 1971; Tinto, 1975), draw on Durkheim's concept of social integration and frame dropout as the result of inadequate academic and social integration. Pascarella and Terenzini (1980) reinforce this, while Bean (1980) emphasizes behavioural and institutional factors, likening dropout to voluntary job turnover. Data-driven research includes econometric analyses (Aina, 2013; Belloc et al., 2011; Contini et al., 2018; Ghignoni, 2017; Gitto et al., 2016) and predictive modelling (Berka and Marek, 2021; Kiss et al., 2019; Perchinunno et al., 2021; Urbina-Nájera and Méndez-Ortega, 2022), employing statistical and machine learning methods to examine academic history, socio-economic status, and demographics. Key findings highlight the impact of family background (Aina, 2013; Belloc et al., 2011; Contini et al., 2018; Ghignoni, 2017; Gitto et al., 2016; Zotti, 2015), age, gender, and geographic origin (Perchinunno et al., 2021), as well as prior performance and first-year results as strong predictors (Belloc et al., 2011; Hoffait and Schyns, 2017; Kehm et al., 2019; Kiss et al., 2019; Perchinunno et al., 2021; Zotti, 2015).

Given this complexity, policies should not address isolated factors but tackle the full range of influences through integrated strategies tailored to diverse student needs.

In this work, we aim to develop a tool enabling a more inclusive and data-driven approach to addressing dropout by considering the diverse factors influencing a student's likelihood of leaving

university before graduation. We leverage data from engineering students at the Politecnico di Milano to identify the most significant predictors of dropout risk. By embedding these predictors into a compact tool, Dropout-DNA, we provide institutions with a practical means to design targeted policies. This tool allows for a nuanced understanding of the dimensions influencing dropout, ensuring preventive interventions are both timely and holistic.

Previous studies at Politecnico di Milano reveal dropout is a critical issue in engineering programs. Nonetheless, several variables regarding students' background and early academic performance can be identified early (Cannistrà et al., 2022; Masci et al., 2024; Pellagatti et al., 2021). In particular, multilevel classification and time-to-event models show that gender, type of previous studies, origin, age and family income are predictive factors, but dropout risk is mostly determined by performance at the beginning of the academic career. The multilevel approach also reveals heterogeneity across engineering courses in terms of dropout risk and timing, net of student characteristics.

Although these studies offer tools to describe and predict dropout, none provides a clear way to profile students. This paper proposes Dropout-DNA, a general framework for profiling and predicting dropout risk. We illustrate the approach using records from engineering programmes at Politecnico di Milano, as a case study that exemplifies the method's application. By means of the Dropout-DNA, we contribute to this literature by proposing a new predictive tool for dropout risk that, in addition, offers a synthetic characterization of student profiles based on the most important dimensions of the phenomenon. Moreover, we investigate the Dropout-DNA heterogeneity across degree programs, studying its course-specific characterization.

Our primary aim is to develop an interpretable, practical tool for predicting dropout risk. This helps institutions gain insights into the varied reasons behind attrition. Our framework combines data exploration with ML algorithms to identify key dropout factors. Dropout-DNA thus contributes a novel way to classify and profile at-risk students.

We validate this approach across multiple programs to demonstrate reproducibility. Considering different stages of the academic journey, the analysis provides a nuanced understanding of dropout risk, supporting targeted and timely policy decisions.

The remainder of the paper is organized as follows: Section 2 introduces the conceptual framework and methodology for developing Dropout-DNA. Section 3 presents the data mining process and exploratory analyses. Section 4 reports the main findings on classification and prediction. Finally, Section 5 discusses implications and offers recommendations for the use of Dropout-DNA.

Methodology

The methodological process follows three main phases.

First, we conduct exploratory data analysis and preprocessing to understand the dataset and address any issues that could affect the subsequent analysis.

In this initial phase, we define the different scenarios of our analysis by partitioning the dataset based on the dropout timing, whether it occurs within the first three semesters or later, and based on the

available features in two different time, those accessible at the moment of student admission and those at the end of the third semester.

Following this,

to identify key factors influencing dropout, we frame the task as supervised classification. After testing several techniques (generalized linear models, gradient boosting, random forest), logistic regression is selected for its simplicity, interpretability, and strong performance in binary classification.

Using this strategy we are able to model the probability of dropout as a function of various predictors and produce an output between 0 and 1 that represents a student's likelihood of dropping out.

To enhance model interpretability and identify key predictors, we use the computationally efficient Permutation Importance algorithm (Altmann et al., 2010). This algorithm measures the decrease in model performance when a feature's values are shuffled, with larger drops indicating more influential features. This allows us to rank attributes by their contribution and highlight the most relevant factors.

The Dropout-DNA tool can subsequently be constructed. We represent this tool as a vector of probabilities, where each value corresponds to the feature-specific marginal probability of student dropout among the significant variables. To achieve this, we analyse the set of students who share a particular value for each predictive feature selected by the logistic regression model. For each of these profiles, we calculate the dropout likelihood by examining the proportion of students within the group who have dropped out. This approach allows us to determine the dropout risk associated with each combination of feature values. Once these probabilities are calculated, each student can be associated with a vector that reflects the dropout probabilities based on their profile, where each element of the vector corresponds to one feature. This personalized representation offers a nuanced assessment of dropout risk, providing institutions with a practical tool to identify at-risk students and implement targeted interventions.

Finally, to assess the validity of the proposed tool, an analysis is conducted to determine whether it could be used as a predictive tool. Specifically, once the the Dropout-DNA has been computed for a new group of students, their computed dropout probabilities are compared with the probability distributions obtained from the original dataset. This approach relies on probability density estimation and allows us to assign each student to one of the two dropout classes, the most probable one.

Data description and preprocessing

The proposed methodology was applied to a dataset comprising 5,575 engineering students enrolled in the bachelor's degree program of Computer Science and Engineering at the Politecnico di Milano between 2010 and 2019.

The dataset includes a vast range of variables, divided into two main groups. The first group covers features related to students' characteristics available at enrolment, including demographic information (age, gender, place of residence), and previous academic background—such as high school type (according to the Italian system) and final grade. It also includes a socioeconomic variable, represented by income bracket categories. Lastly, this group contains data from the admission test, including the score and whether the student was assigned Obligatory Formative Activities (OFAs). OFAs are mandatory credits for students scoring below 60/100 and restrict exam access until completed.

The second group of variables captures information about the exams taken by each student throughout their university career, organized by semester. At Politecnico di Milano, exams can typically be taken at the end of each semester, with an additional session available after the summer break. This dataset includes detailed records of students' academic performance, such as the number of exams and University Credits (CFUs) taken and passed, and the average grade.

Finally, the dataset includes the target variable, a binary feature indicating whether the student dropped out before completing the degree. Additionally, the dataset provides information on when the dropout occurred, enabling more detailed temporal analyses of the phenomenon.

Data cleaning

The data cleaning process is crucial to ensure affordability in classification analysis (Alyahyan and Düştegör, 2020). To address high feature dimensionality, we selected a reduced set of variables. We began with a preliminary correlation analysis to identify highly correlated characteristics, aiming to improve model robustness and interpretability. We also drew on existing literature to enhance predictive power and reduce overfitting.

To summarize exam performance compactly and interpretably, we used a Weighted Performance Index (WPI), defined as:

$$WPI = GPA \times ln(1 + CFU) \quad (1)$$

Here, GPA denotes the weighted average of exam grades based on course credits (CFU), thus measuring performance quality, while CFUs earned represent workload quantity. Calculated over the first three semesters, WPI integrates both aspects into a single metric of academic success.

This process yielded a set of key variables summarized in Table 1, chosen for a balance between interpretability and predictive power. The first set of variables describes students' demographic and socioeconomic characteristics. We consider a variable for origins, which classifies students based on their nationality and residency status, distinguishing between local students that can be commuter, living away from home, local to the university area or international students. Another key socioeconomic indicators are the income bracket, which serves as a proxy for the financial background of students, gender, and age at admission, which may reflect differences in educational pathways. The second group of features captures students' pre-university academic background. This includes the type of high school attended, which helps account for differences in educational experiences prior to university. We also consider academic performance indicators such as the final high school grade and the university admission score. Additionally, we include information on whether the student was assigned any OFAs to reflect possible gaps in preparation for university-level coursework. Finally, as already mentioned, we have included three indicators related to the early stages of the students' academic career, meaning the WPI of the first, second and third semester. The last key variable in our feature set is, evidently, our target variable, the one related to dropout.

Data partitioning

To provide a more nuanced understanding of the dropout phenomenon, we performed four different analyses by distinguishing between two types of dropout timing and two sets of features.

Students in the dataset drop out at different stages in their academic careers, ranging from a few months after enrolment to several years later.

To account for the variability of dropout timing, we categorized dropouts into two cases: *early dropouts* and *late dropouts*. Dropouts that occur within the first three semesters of a career are classified as *early dropouts*, whereas *late dropouts* are those that take place after the third semester. This distinction is crucial, as the underlying motivations and contributing factors for these two groups may differ significantly and conducting separate analyses enables us to uncover these differences more effectively.

Moreover, to support the design of more tailored interventions, we considered two sets of features, as summarized in Table 1. We collected data at enrolment and for each semester up to the third. Then, depending on the type of dropout we aim to predict and the timing of the prediction, we propose four different approaches. This approach allows us to identify students' specific needs at critical points in their academic careers, enhancing the policymakers to act at different moments of time to improve the effectiveness of measures aimed at reducing attrition and supporting student success.

Table 1. Set of variables considered for the analysis.

Attribute	Type	Category	Possible values	Descriptive statistics
Origins	Cat	At admission	Commuter	69.6%
			Milanese	21.5%
			Offsite	6.1%
			Foreigner	2.8%
Income bracket	Cat	At admission	Highest bracket	13.2%
			High bracket	39.6%
			Low bracket	34.2%
			DSU (students who receive scholarship)	13.0%
Age at admission	Num	At admission	From 16 to 58	19.35 (IQR 18.0 - 19.0)
Gender	Cat	At admission	1: male	89.9%
			0: female	10.1%
Highschool type	Cat	At admission	Scientific	63.5 %
			Technical	21.8%
			Classic	4.2%
			Other	10.5%
Highschool final grade	Num	At admission	From 60 to 100	82.00 (IQR 72.0 - 93.0)
Admission score	Num	At admission	From 60 to 100	83.32 (IQR 83.30 – 90.24)
OFA at admission	Cat	At admission	0: No	90.3%
			1: Yes	9.7%
WPI of first semester	Num	End of first semester	From 0 to 47.73	20.47 (IQR 0.0 - 33.53)

Attribute	Type	Category	Possible values	Descriptive statistics
WPI of second semester	Num	End of second semester	From 0 to 49.86	22.19 (IQR 0.0 - 36.44)
WPI of third semester	Num	End of third semester	From 0 to 47.58	15.44 (IQR 0.0 - 31.07)
Dropout	Cat	Target	0: No	68%
			1: Yes	32%

Consequently, the analysis is structured around four distinct cases, each corresponding to a unique combination of dropout timing and feature set:

- **Case A – Outcome: Early Dropout; Features: Early Indicators**: This case leverages data available at the time of enrolment to predict early dropouts, focusing on demographic, high school, and admission test variables. It should be noted that, prior to the balancing step, this case includes all students: 4,668 who remained enrolled for at least the first three semesters and 907 who dropped out during that period.

- **Case B – Outcome: Early Dropout; Features: Academic Progress**: This extends the previous case by incorporating information about academic performance, such as early academic performance. In particular, here we consider the WPI calculated from the first semester only in order to ensure that the temporal scope remains aligned with the timing of early dropouts. As above-mentioned the students considered in this case are all those in the dataset.

- **Case C – Outcome: Late Dropout; Features: Early Indicators**: Here, the analysis seeks to predict late dropouts using only the information available at the time of enrolment. The students included in this case are those who persisted until the third semester thus excluding those who early-dropped out. In particular there are 3,298 students that don't drop out and 1,370 students that drop out.

- **Case D – Outcome: Late Dropout; Features: Academic Trajectory**: In this case, both enrolment data and cumulative academic performance up to the third semester (WPI) are utilized to identify late dropouts. As in the previous case, students who dropped out within the early three semesters are not included in the analysis of this case.

For clarity, a summary table (see Table 2) is provided, showing which features are considered in the four listed cases.

Table 2. Early dropout: original and balanced ratio of dropout classes.

Feature	Case A	Case B	Case C	Case D
Origins	X	X	X	X
Income bracket	X	X	X	X
Age at admission	X	X	X	X
Gender	X	X	X	X
Highschool type	X	X	X	X

Feature	Case A	Case B	Case C	Case D
Highschool final grade	X	X	X	X
Admission score	X	X	X	X
OFA at admission	X	X	X	X
WPI of first semester		X		X
WPI of second semester				X
WPI of third semester				X

Data balancing

As shown in Table 3, when considering early dropout we have a highly unbalanced dataset, differently from what can be seen from the late dropout case (see Table 4). This may impact in a negative way the performance of our classification model (Alyahyan and Düştegör, 2020). To balance our datasets, we under-sampled the majority class — the dropout 0 class - applying a random undersampling technique. This method randomly removes samples to balance the class distribution without introducing synthetic data.

Although random undersampling does not guarantee exact preservation of the original feature distributions, we ensured that the sampling strategy maintained a representative subset of the original data. The resulting ratio between the two classes is shown in Table 3.

Table 3. Early dropout: original and balanced ratio of dropout classes.

Class	Size (Original)	Ratio (Original)	Size (Balanced)	Ratio (Balanced)
Dropout 0	4668	84%	2267	60%
Dropout 1	907	16%	905	40%

Table 4. Late dropout: ratio of dropout classes.

Class	Size (Original)	Ratio (Original)
Dropout 0	3298	60%
Dropout 1	1367	40%

Results

Based on the resulting datasets described in the previous section, we employ the tool of logistic regression in order to build a supervised classification model. For each of the four cases described in the previous section, we split the datasets into a training and a test set (75% - 25%).

The model was estimated on the training set, comprising 75% of the dataset, and produced the results shown in Table 5. While we report Case B in detail, for the sake of conciseness, the detailed results for all cases (A, C, and D) are not reported here, but yield qualitatively consistent outcomes, with variations mainly in the relative strength of predictors depending on the available information at each

stage. The model demonstrates a good explanatory power, as evidenced by the Pseudo R^2 value of 0.2879. Additionally, the model is statistically significant with a likelihood ratio p-value of less than 0.001. Several variables exhibit strong statistical significance. For instance, the variable representing the students' age has a coefficient of -0.0771 with a p-value of less than 0.001, suggesting that older admission age is associated with a lower dropout risk. Similarly, the income variable shows a positive coefficient of 0.0444 with a p-value of 0.004, indicating that higher income slightly increases the risk of dropout. The variable related to the high school grade has a coefficient of 0.0080 and a p-value of 0.046, suggesting a small positive correlation between the high school performance and dropout risk. Lastly, the WPI of the first semester is highly significant with a negative coefficient of -0.0984 and a p-value of less than 0.001. In contrast, gender, high school type, geographic origin, and admission score were not statistically significant. For categorical variables, the reported coefficients refer to comparisons with the respective reference categories.

Table 5. Summary of results of Logistic Regression – Case B. Coefficients for categorical variables are relative to the following reference categories: Scientific (high school type), female (gender), and Milanese (origins).

Variabile	**Coeff.**	**Std. Err.**	**z**	**p-value**
Intercept (const)	0.5613	0.296	1.900	0.057
Highschool type:Technical	0.150	0.134	1.12	0.262
Highschool type:Classical	0.0832	0.288	0.289	0.772
Highschool type:Other	0.1694	0.283	0.599	0.549
Income bracket	0.0444	0.016	2.851	**0.004**
Admission score	0.00003	0.0001	0.291	0.771
Age at admission	-0.0771	0.019	-4.076	**0.000**
Gender:male	-0.2700	0.195	-1.383	0.167
Origins:Commuter	0.1560	0.136	1.148	0.251
Origins:Offsite	-0.0038	0.259	-0.015	0.988
Origins:Foreigner	0.2882	0.469	0.614	0.539
OFA at admission	0.0964	0.186	0.519	0.603
Highschool final grade	0.0080	0.004	1.997	**0.046**
WPI of first semester	-0.0984	0.004	-22.586	**0.000**

To evaluate the ranking importance of the features considered in the analysis, we calculated the permutation importance. Permutation importance provides insight into the relative importance of each feature by measuring how shuffling each feature's values affects model performance. A larger drop indicates greater importance, while a smaller drop suggests lower impact. Note that dummy variables originating from the same categorical feature are permuted together to assess their overall contribution.

As shown in Figure 1, WPI of the first semester is the most important variable, significantly impacting model performance. High school grade, income, and age at admission also contribute meaningfully.

Figure 1. Permutation importance results - Case B

Another important aspect of evaluating a classification model is to understand its performance. Given that TP, FP and FN are respectively the number of true positives, false positives and false negatives

obtained after applying the learned classifier on the test set, we consider the standard performance metrics as follows:

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2PR}{P + R}, \tag{2}$$

The resulting performance of the classification model for Case B is shown in Table 6.

Table 6. Logistic regression performance - Case B

Class	Precision	Recall	F1-Score	Support
Dropout 0	87%	83%	85%	567
Dropout 1	61%	68%	64%	226

The model performs well in predicting students who do not drop out (class of Dropout 0), with high precision (87%) and recall (83%). For dropouts (class of Dropout 1), the model achieves a precision of 61% and a recall of 68%, indicating a fairly good ability to identify at-risk students. Comparable performance was observed across the other cases (A, C, and D), reinforcing the robustness of the results.

After evaluating both the statistical significance of features in the logistic regression models and their importance as indicated by permutation importance, we selected the features to be included in the Dropout-DNA. In particular, we have decided to consider two distinct scenarios, early indicators and academic progress. This choice ensures consistency in the selection of predictive variables across different dropout timings, which facilitates the implementation of predictive models in real-world interventions. From a practical perspective, to counter dropout the primarily concern is to identify when and on which factors to intervene, rather than distinguishing between early and late dropout per se. By focusing on these two scenarios, we provide a clearer framework for tailoring interventions based on the available data at different stages of a student's academic journey. In particular, for the early indicators case, we select the features related to the high school grade, the admission score, the age at admission, the origins and the income of the students. For the academic progress case, we expand this selection to include a feature associated with academic performance, in particular, the WPI related to the first semester of students' career that appears in both cases (Cases B and D) at the front of the features ranking.

Dropout-DNA computation

For our objective of constructing the Dropout-DNA, we consider the attributes selected in the previous section. To construct a tool that indicates the predisposition to dropout, we need to quantify the relationship between each of the selected features and the target variable. We consider the two subsets of students, namely the positive individuals P, students who decided to drop out of school, and negative individuals N, students who continued their course of study. Furthermore, we define each selected feature i, with $i = 1, \dots, 5$ for the case of early dropout and $i = 1, \dots, 6$ for the case of late dropout, and each possible value $j = 1, \dots, m_i$ that attribute i can take, where m_i represents the number of distinct values attribute i can assume. For categorical features, m_i corresponds to the number of unique categories. For numerical features, we discretise the values into a finite set of intervals, so that each value can be treated analogously to a category. In this way, we can define the group of individuals whose $i - th$ attribute takes the $j - th$ possible value, denoted as $A_{i,j}$, and the likelihood with which these individuals belong to the set of positive individuals as follows:

$$l_{i,j} = \frac{\#(P \cap A_{i,j})}{\#A_{i,j}} \quad (3)$$

with $\#(\cdot)$ being the cardinality of a set. Each student is assigned a unique Dropout-DNA from their feature values. Its vector format allows effective visualisation via spider plots. Figure 2a-2d show two random examples for each class: Dropout-DNA of students who discontinue their studies (taking as an example the late dropout scenario, Figure 2b and Figure 2d) and one of students who persist in their academic journey (Figure 2a and Figure 2c).

Figure 2a. Spider plot visualizing the Dropout-DNA profiles for students who dropped out versus those who persisted, based on early indicators and academic progress. *Features: Early Indicators - Negative student DNA*

Figure 2b. Spider plot visualizing the Dropout-DNA profiles for students who dropped out versus those who persisted, based on early indicators and academic progress. *Features: Early Indicators - Positive student DNA*

Figure 2c. Spider plot visualizing the Dropout-DNA profiles for students who dropped out versus those who persisted, based on early indicators and academic progress. *Features: Academic Progress - Negative student DNA*

Figure 2d. Spider plot visualizing the Dropout-DNA profiles for students who dropped out versus those who persisted, based on early indicators and academic progress. *Features: Academic Progress - Positive student DNA*

The spider plots reveal clear differences between the DNAs of dropout and non-dropout students, with dropout students consistently showing smaller plot areas. This distinction is further emphasized in Figure 3a, Figure 3b, Figure 3c and Figure 3d where the area distributions for all four cases show a clear separation between the two classes.

Figure 3a. Dropout-DNA distribution: positive and negative students classes for the four cases. Case A.

Figure 3b. Dropout-DNA distribution: positive and negative students classes for the four cases. Case B.

Figure 3c. Dropout-DNA distribution: positive and negative students classes for the four cases. Case C.

Figure 3d. Dropout-DNA distribution: positive and negative students classes for the four cases. Case D.

Dropout-DNA of different courses of study

To asses the robustness and generalizability of the proposed methodology, the same procedure was fully applied to other two datasets related to the Mechanical Engineering and Chemical Engineering programs of study. This approach enables a comparison across degree programs in terms of model structure and performance. Furthermore, this extension allows the inclusion of the datasets potentially composed of students with different characteristics. In particular, in the analysis conducted in the previous section, gender never emerged as a relevant feature. Since this outcome differs from what is commonly reported in the literature, we decided to include two datasets with different balances of female and male students. In particular, the Computer Science and Engineering datasets comprised 90% of male students and 10% of female students. To further explore this aspect, we selected one program with a more evenly balanced proportion (Chemical Engineering) that comprised 55% of male students and 45% of female students. Moreover, we decided to consider also an even more unbalanced gender ratio with the Mechanical Engineering program that featured 93% of male students and 7% of

female students. Including these diverse datasets helps assess whether such demographic variations impact the predictive features identified by the methodology.

The two datasets comprised 5376 students for the Mechanical Engineering dataset and 1736 students for the Chemical Engineering dataset, and they included information from the years 2010 and 2019 as the previous dataset. Furthermore, the datasets included the same range of variables, therefore, the initial steps of data pre-processing and data cleaning led to the same considerations. In particular, to ensure a fair comparison, we considered the same cases considered in the previous analysis, namely Case A, Case B, Case C and Case D with the same selection of features as summarised in Table 2.

Table 7. Mechanical Engineering - Early dropout: original and balanced ratio of dropout classes.

Class	Size (Original)	Ratio (Original)	Size (Balanced)	Ratio (Balanced)
Dropout 0	4697	87%	1689	63%
Dropout 1	679	13%	617	37%

Table 8. Mechanical Engineering - Late dropout: original and balanced ratio of dropout classes.

Class	Size (Original)	Ratio (Original)	Size (Balanced)	Ratio (Balanced)
Dropout 0	3775	80%	2305	62%
Dropout 1	922	20%	885	38%

Table 9. Chemical Engineering - Early dropout: original and balanced ratio of dropout classes.

Class	Size (Original)	Ratio (Original)	Size (Balanced)	Ratio (Balanced)
Dropout 0	1495	86%	602	60%
Dropout 1	241	14%	239	40%

Table 10. Chemical Engineering - Late dropout: original and balanced ratio of dropout classes.

Class	Size (Original)	Ratio (Original)	Size (Balanced)	Ratio (Balanced)
Dropout 0	1279	86%	540	60%
Dropout 1	216	14%	214	40%

As Table 7, Table 8, Table 9 and Table 10 show, there are differences in the distribution of the two classes with respect to the Computer Science and Engineering case (see Table 3 and Table 4). In particular, for both new datasets, the ratio between Dropout 0 and Dropout 1 class for both the early dropout and late dropout case are more unbalanced. Therefore, to once again achieve comparability between analyses, we employed the balancing approach as illustrated previously for all cases.

The same classification pipeline was applied to both datasets.

Coefficient and p-value analysis reveal both similarities and significant differences across the various degree programs. The WPI variable proves to be significant in all study programs whenever it is included in the model. Other variables that are consistently significant across all three programs are highschool final grade, age at admission, and origins. In contrast, income and OFA at admission are significant only for Computer Engineering and Mechanical Engineering, but not for Chemical Engineering. The gender variable, on the other hand, is significant in only a few cases.

Regarding the permutation importance analysis, several of the aforementioned observations are confirmed. In particular, the WPI consistently ranks highly in all programs. Highschool final grade also shows high importance, while age at admission generally ranks moderately. The income bracket is confirmed to be less important in Chemical Engineering, where the origins variable plays a more relevant role. Finally, the OFA at admission variable shows lower importance in both Chemical and Mechanical Engineering. Gender remains a consistently low-importance variable across all programs.

As regards the performance metrics, similar to the previous case, the performances result to be particularly satisfactory for Case B and Case D with a slight worsening for the performances of Chemical Engineering's Class of Dropout 1 in the Case C. The pattern between Case A and B and between Case C and D is again notable, with the performances consistently improving in the latter.

Figure 4a. Mechanical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. *Case A.*

Figure 4b. Mechanical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. *Case B.*

Figure 4c. Mechanical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. *Case C.*

Figure 4d. Mechanical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. *Case D.*

Figure 5a. Chemical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. Case A.

Figure 5b. Chemical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. Case B.

Figure 5c. Chemical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. Case C.

Figure 5d. Chemical Engineering - Dropout-DNA distribution: positive and negative students classes for the four cases. Case D.

For the sake of a consistent comparison, and considering that the feature rankings derived from the permutation importance don't show substantial differences, the feature selection for constructing the Dropout-DNA of Mechanical and Chemical Engineering students remains the same as previously selected. Following the methodology for the construction of the Dropout-DNA, as in the previous case, we observe that the distributions of the corresponding areas for the two dropout classes are well-separated (see Figure 4a, Figure 4b, Figure 4c, Figure 4d and Figure 5a, Figure 5b, Figure 5c, Figure 5d).

Prediction on different courses of study

To assess the possibility of using this tool as a predictive instrument, we extended our evaluation by using the results obtained from the Computer Science and Engineering dataset as a reference to classify enrolled students from the Mechanical Engineering and Chemical Engineering programs.

In particular, the classification was performed by comparing the estimated probability densities at the area value observed for each new student. The two densities, p_1 and p_0, were estimated using kernel density estimation based on the Dropout-DNA areas of students in the Computer Science and Engineering dataset divided in students who dropout (D_1) and studetns who don't dropout (D_0). For each new student v enrolled in Mechanical or Chemical Engineering, we first computed their area value $area_v$ using the same set of features used to build the reference Dropout-DNA. Then, we evaluated how compatible this area value was with each of the two estimated distributions. The student was assigned to the class for which their area value corresponded to a higher probability density. More formally, this corresponds to assigning the student to the class in which their area value falls in a higher quantile of the corresponding density, as summarized by the following decision rule:

$$PredictedClass_v = \begin{cases} D_1 & if\ p_1(area_v) > p_0(area_v) \\ D_0 & otherwise \end{cases} \quad (4)$$

where $p_1(\text{area}_v)$ and $p_0(\text{area}_v)$ represent the estimated probability densities for the two dropout classes, evaluated at the area value computed for student v.

To evaluate the predictive performance, the predicted labels were compared against the true labels using the standard performance metrics in (2). A representative subset of results is presented in Table 11 and Table 12, which report the outcomes when considering Case B.

For the Mechanical Engineering dataset, the results shows that Case B achieves promising values, with an F1-score of 77% for Dropout 0 and 50% for Dropout 1.

A similar trend is observed in the Chemical Engineering dataset, where Case B also performs well, yielding an F1-score of 78% for Dropout 0 and 51% for Dropout 1.

The corresponding confusion matrices are shown in Figure 6 and Figure 7. In both case, the model demonstrates moderately balanced predictive performance, with a notable ability to correctly identify students at risk of dropout (Dropout 1), reflected in relatively high true positive rates and a manageable number of false negatives.

Table 11. Prediction: performance indexes over the classification of Mechanical Engineering students - Case B

Class	Precision	Recall	F1-Score	Support
Dropout 0	80%	74%	77%	1689
Dropout 1	46%	54%	50%	617

Figure 6. Prediction: confusion matrix over the classification of Mechanical Engineering students - Case B

Table 12. Prediction: performances indexes over the classification of Chemical Engineering students - Case B

Class	Precision	Recall	F1-Score	Support
Dropout 0	81%	75%	78%	602

Class	Precision	Recall	F1-Score	Support
Dropout 1	47%	55%	51%	239

Figure 7. Prediction: confusion matrix over the classification of Chemical Engineering students - Case B

Prediction of the same course of study

To assess the predictive accuracy of the model when applied to students in the same degree programme, we split the original dataset of Computer Science and Engineering students into two subsets: a training subset, which was used to build the classification model and the Dropout-DNAs, and a test subset, on which predictions were made. This cross-validation procedure allows us to assess the performance of the model without introducing bias from external datasets.

As with the other analyses, we considered the four separate cases (Cases A, B, C and D) to maintain methodological consistency. The classification was performed using the probabilities of belonging to the two dropout classes estimated by kernel density estimation (KDE) models based on the area distributions calculated in the training dataset. The decision rule used is the same as that described in (4).

The performance assessment metrics shown in Table 13 indicate that the model applied to the Computer Science and Engineering dataset performs comparably to the results obtained for the Mechanical and Chemical Engineering datasets in Case B. Specifically, Case B achieves an F1-score of 75% for Dropout 0 and 50% for Dropout 1, confirming its predictive effectiveness across different courses of study. However, However, across the other cases we observed improved performance compared to Case B. A visual inspection of the confusion matrices (Figure 8) further confirms these findings, revealing a reduction in false positives compared to the scenarios involving Mechanical and Chemical Engineering. This trend is consistently observed across the remaining cases.

Finally, the predictive performance observed in Table 13 demonstrates strong consistency with the results from the logistic regression models presented in the earlier analysis. Specifically, the F1-scores obtained in the prediction phase for both Dropout 0 and Dropout 1 classes align closely with those reported in Table 6, as well as in the corresponding results for the other cases.

Table 13. Prediction: performances indexes over the classification of Computer Science and Engineering students - Case B

Class	Precision	Recall	F1-Score	Support
Dropout 0	80%	71%	75%	226
Dropout 1	44%	57%	50%	91

Figure 8. Prediction: confusion matrix over the classification of Computer Science and Engineering students - Case B

Discussion, implications and concluding remarks

The classification results show valuable insights into the dropout phenomenon and present more than satisfactory performance. In particular, we can say that the models relying solely on early indicators (Cases A and C) are generally less predictive than those that incorporate information about academic progress (Cases B and D). This is evident in the notable improvement in F1-scores for both dropout classes when dynamic and longitudinal data such as the WPI are included. Furthermore, the classification of the Dropout 1 class consistently shows lower performance compared to the Dropout 0 class, a gap that narrows when the information about academic performance is included. This may be partially attributed to the balancing of the dataset used for Case A as the use of undersampling techniques may have influenced the classification results.

Regarding the classification performed using the Mechanical and Chemical Engineering datasets, some differences were observed, particularly in the case of Chemical Engineering. This may be partly attributed to the differing ratios between early and late dropouts. In fact, while in the cases of Computer Science and Engineering and Mechanical Engineering the majority of dropouts belonged to the late dropout category, in the case of Chemical Engineering early dropouts were more prevalent.

Beyond the standard classification metrics, an important validation of our approach comes from the Dropout-DNA itself. The separation of the distribution of the areas defined by the spider plots of the students' DNAs of all courses of study under analysis tells us that the model captures the key dropout predictors. The clear difference between the size of the areas, bigger for the students who dropout and smaller for the students who don't, further validates the ranking of the features obtained by the permutation importance algorithm. s a result, our tool offers a novel, visual way to represent dropout risk. The visual aspect is particularly valuable, as it transforms complex multidimensional data into an interpretable format that stakeholders can easily understand. This contribution of our work enhances the practical utility of dropout risk assessment.

The predictor's results also offer several insights into the potential of the Dropout-DNA tool and show the model's reproducibility. While predictions for Mechanical and Chemical Engineering perform well, the tool works best when applied to the dataset of the degree program it was built on. This is supported by both quantitative metrics and visual evidence from the confusion matrices, which illustrate the areas where the model underperforms. Notably, the reduction of the false negatives implies a better capacity of the model to identify the students at risk. This may be due to the phenomenon's complexity, as some predictors could not be consistent across different academic departments. In fact, such complexity may result from both contextual factors and differing student characteristics across degree programs. This is particularly evident in the case of the Chemical Engineering dataset, where the prediction performance was slightly lower than in Mechanical Engineering. This aligns with our findings showing greater differences between Chemical and Computer Science students than between Mechanical and Computer Science students.

These findings not only highlight the importance of contextualising predictive models but also reinforce the need for adaptable, yet interpretable tools in dropout risk assessment. In this context, this study introduces the Dropout-DNA, a novel data-driven tool that can be used to profile students and predict early university dropout risk. This is a significant contribution, offering a powerful approach to understanding and addressing student retention. Our findings indicate that early indicators matter, but academic progress significantly enhances predictive performance.

Moreover, the defined predictive approach can be generalized to various engineering courses, though performance improves when tailored to the specific degree program. This underscores the need for further exploration into how the Dropout-DNA may adapt across different degree courses with different intrinsic characteristics, as dropout patterns vary.

These results also have important practical implications. By enabling the early identification of at-risk students through the detailed, quantifiable representation offered by the Dropout-DNA, universities can intervene with tailored policies. Such policies can address individual dropout risk factors, allowing more personalized support. In future work, we aim to explore the temporal dimension of the Dropout-DNA more explicitly by investigating how student dropout risk profiles evolve throughout different stages of the academic journey. This would allow for dynamic monitoring and the potential to design time-sensitive interventions aligned with students' changing academic trajectories. In particular, as in our previous contributions (Villa, Breschi, and Tanelli, 2023) and (Villa, Breschi, Ravazzi, et al., 2023), we aim to apply a quantitative framework based on optimal control theory and opinion dynamics to support policy design that prevents dropout.

Acknowledgements

The present research is part of the activities of "Dipartimento di Eccellenza 2023-2027"

References

Aina, C. (2013). Parental background and university dropout in italy. *Higher Education*, *65*(4), 437–456. https://doi.org/10.1007/s10734-012-9554-z

Altmann, A., Tolosi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics (Oxford, England)*, *26*, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

Alyahyan, E., and Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 3. https://doi.org/10.1186/s41239-020-0177-7

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187. https://www.jstor.org/stable/40195329

Belloc, F., Maruotti, A., and 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. https://doi.org/10.1080/02664763.2010.545373

Berka, P., and Marek, L. (2021). Bachelor's degree student dropouts: Who tend to stay and who tend to leave? *Studies in Educational Evaluation*, *70*, 100999. https://doi.org/10.1016/j.stueduc.2021.100999

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2022). Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, *47*(9), 1935–1956.

Contini, D., Cugnata, F., and Scagni, A. (2018). Social selection in higher education. Enrolment, dropout and timely degree attainment in italy. *Higher Education*, *75*(5), 785–808. https://doi.org/10.1007/s10734-017-0170-9

Ghignoni, E. (2017). Family background and university dropouts during the crisis: The case of italy. *Higher Education*, *73*(1), 127–151. https://doi.org/10.1007/s10734-016-0004-1

Gitto, L., Minervini, L. F., and Monaco, L. (2016). University dropouts in italy: Are supply side characteristics part of the problem? *Economic Analysis and Policy*, *49*, 108–116. https://doi.org/10.1016/j.eap.2015.12.004

Hoffait, A.-S., and Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems*, *101*, 1–11. https://doi.org/10.1016/j.dss.2017.05.003

Kehm, B. M., Larsen, M. R., and Sommersel, H. B. (2019). Student dropout from universities in europe: A review of empirical literature. *Hungarian Educational Research Journal*, *9*(2), 147–164. https://doi.org/10.1556/063.9.2019.1.18

Kiss, B., Nagy, M., Molontay, R., and Csabay, B. (2019). Predicting dropout using high school and first-semester academic achievement measures. *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158

Lorenzo-Quiles, O., Galdón-López, S., and Lendínez-Turón, A. (2023). Factors contributing to university dropout: A review. *Frontiers in Education*, *8*, 1159864. https://doi.org/10.3389/feduc.2023.1159864

Masci, C., Cannistrà, M., and Mussida, P. (2024). Modelling time-to-dropout via shared frailty cox models. A trade-off between accurate and early predictions. *Studies in Higher Education*, *49*(4), 763–781.

Ministero dell'Università e della Ricerca. (2023, May 22). *Università: MUR, rafforziamo orientamento e diritto allo studio per contrastare abbandono*. https://www.mur.gov.it/it/news/lunedi-22052023/universita-mur-rafforziamo-orientamento-e-diritto-allo-studio-contrastare

Pascarella, E. T., and Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, *51*(1), 60–75. https://doi.org/10.2307/1981125

Pellagatti, M., Masci, C., Ieva, F., and Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(3), 241–257.

Perchinunno, P., Bilancia, M., and Vitale, D. (2021). A statistical analysis of factors affecting higher education dropouts. *Social Indicators Research*, *156*(2), 341–362. https://doi.org/10.1007/s11205-019-02249-y

Skrbinjek, V., Lesjak, D., and Dermol, V. (2024, May). Higher education dropout: A literature review. *Proceedings of the Management, Knowledge and Learning International Conference (MakeLearn) and Technology, Innovation and Industrial Management (TIIM)*.

Spady, W. G. (1971). Dropouts from higher education: Toward an empirical model. *Interchange*, *2*(3), 38–62. https://doi.org/10.1007/BF02282469

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125. https://doi.org/10.2307/1170024

Urbina-Nájera, A. B., and Méndez-Ortega, L. A. (2022). Predictive model for taking decision to prevent university dropout. *International Journal of Interactive Multimedia and Artificial Intelligence*, *7*(4), 205. https://doi.org/10.9781/ijimai.2022.01.006

Villa, E., Breschi, V., Ravazzi, C., Dabbene, F., and Tanelli, M. (2023). Fostering the use of sharing mobility solutions via control-oriented policy design. *IFAC-PapersOnLine*, *56-2*, 1–6.

Villa, E., Breschi, V., and Tanelli, M. (2023). Fair-MPC: A control-oriented framework for socially just decision-making. *Submitted*. https://arxiv.org/abs/2312.05554

Zotti, R. (2015). Should i stay or should i go? Dropping out from university: An empirical analysis of students' performances. In G. Coppola and N. O'Higgins (Eds.), *Youth and the crisis* (1st ed., p. 18). Routledge.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

56/2025 Tonini, A.; Bui-Thanh, T.; Regazzoni, F.; Dede', L; Quarteroni, A.
Improvements on uncertainty quantification with variational autoencoders

Tonini, A.; Bui-Thanh, T.; Regazzoni, F.; Dede', L; Quarteroni, A.
Improvements on uncertainty quantification with variational autoencoders

55/2025 Gimenez Zapiola, A.; Boselli, A.; Menafoglio, A.; Vantini, S.
Hyper-spectral Unmixing algorithms for remote compositional surface mapping: a review of the state of the art

Gimenez Zapiola, A.; Boselli, A.; Menafoglio, A.; Vantini, S.

54/2025 Tomasetto, M.; Williams, J.P.; Braghin, F.; Manzoni, A.; Kutz, J.N.
Reduced order modeling with shallow recurrent decoder networks

53/2025 Zecchi, A. A.; Sanavio, C.; Perotto, S.; Succi, S.
Telescopic quantum simulation of the advection-diffusion-reaction dynamics

51/2025 Tomasetto, M.; Braghin, F., Manzoni, A.
Latent feedback control of distributed systems in multiple scenarios through deep learning-based reduced order models

50/2025 Bonetti, S.; Botti, M.; Antonietti, P.F.
Conforming and discontinuous discretizations of non-isothermal Darcy–Forchheimer flows

49/2025 Zanin, A.; Pagani, S.; Corti, M.; Crepaldi, V.; Di Fede, G.; Antonietti, P.F.; the ADNI
Predicting Alzheimer's Disease Progression from Sparse Multimodal Data by NeuralODE Models

48/2025 Temellini, E.; Ballarin, F.; Chacon Rebollo, T.; Perotto, S.
On the inf-sup condition for Hierarchical Model reduction of the Stokes problem