Balancing Performance and Explainability in Academic Dropout Prediction

Andrea Zanellati, Stefano Pio Zingaro, and Maurizio Gabbrielli

Abstract—**Academic dropout remains a significant challenge for education systems, necessitating rigorous analysis and targeted interventions. This study employs machine learning techniques, specifically random forest (RF) and feature tokenizer transformer (FTT), to predict academic attrition. Utilizing a comprehensive dataset of over 40 000 students from an Italian university, the research incorporates a range of variables, including demographic information, prior educational metrics, and real-time academic performance indicators. We present a nuanced comparative evaluation of the RF and FTT models, highlighting their predictive accuracy and interpretative capabilities. Our empirical results demonstrate the effectiveness of machine learning in managing student attrition, with FTT models outperforming RF models in terms of predictive accuracy and achieving a sensitivity rate of 81%. Significantly, the inclusion of historical academic data enhances the models' ability to identify students at increased risk of dropping out. Furthermore, we apply advanced explanatory techniques, such as shapley additive explanations, to investigate the discriminative power of these models across different student profiles. This provides valuable insights into the key variables influencing dropout risk, contributing to a more holistic understanding of the issue. In addition, we conduct a fairness analysis to ensure the ethical robustness of our predictive models, making them not only effective but also equitable tools.**

Index Terms—**Academic dropout, educational data mining, explainable artificial intelligence (XAI), informed machine learning (ML).**

I. Introduction

DROPOUT is a critical issue in the field of education, with significant consequences for individuals and society as a whole. The complexity and importance of this phenomenon have prompted research efforts since the 70s [1]. In recent years, factors such as the increased accessibility of higher education, the globalization of institutions, digitization, and data-driven practices have reinvigorated the focus on addressing this issue. In response, researchers and educators are exploring innovative approaches, including the integration of artificial intelligence (AI) techniques [2] into educational decision-making processes. AI, encompassing a wide range of techniques such as machine learning (ML), has demonstrated remarkable effectiveness in predictive and diagnostic tasks in a variety of domains.

The integration of AI in education offers direct improvements to the education system. AI enables proactive strategies to anticipate risks and implement interventions. By analyzing student performance, demographics, and socioeconomic factors, educators can tailor learning experiences and allocate resources effectively. This data-driven approach also helps policymakers make informed decisions about educational interventions based on population density, accessibility, and infrastructure. In addition, AI algorithms can recommend professional development courses by analyzing skills gaps, employee performance, and industry trends.

Motivated by the transformative potential of AI, our research aims to develop a robust AI-based tool to address academic attrition. We focus specifically on the outcomes of first-year students, using ML techniques to analyze real data from a prestigious Italian university in an in-person learning setting. We define dropout as a situation where a student does not re-enroll in the same study program for the following academic year. Therefore, the dropout target is always assessed after 12 months of enrollment.

To ensure accurate analysis, we use state-of-the-art decision-tree-based techniques; specifically, we establish our baseline as random forest (RF), known for its predictive performance and explanatory power in previous case studies of academic dropout prediction [3]. In addition, we aim to go beyond conventional modeling approaches by exploring the potential of deep neural networks (DNNs). Our aim is to assess whether their implementation can improve the predictive performance by exploiting their ability to account for nonlinear correlations. Among several solutions, we rely on the feature tokenizer transformer (FTT) approach, a deep architecture that combines tokenization for tabular data representation [4] and an attention mechanism for classification [5], trained end-to-end. Using the transformer technique, we explore flexible strategies for handling the categorical data prevalent in our dataset to adequately represent the findings. To the best of our knowledge, there are no other studies in the literature applying FTT to address the prediction of academic dropout risk. As a result, we address the following research question (RQ).

RQ1: To what extent does the use of FTT improve predictive models of student dropout compared to state-of-the-art techniques?

Manuscript received 23 June 2023; revised 31 October 2023, 18 March 2024, and 25 June 2024; accepted 1 July 2024. Date of publication 26 July 2024; date of current version 3 October 2024. This work was supported in part by Future AI Research (FAIR) PE01, SPOKE 8 on PERVASIVE AI funded by the Italian National Recovery and Resilience Plan (NRRP). *(Corresponding author: Stefano Pio Zingaro.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Bologna.

The authors are with the Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy (e-mail: andrea.zanellati2@unibo.it; stefano.zingaro@unibo.it; maurizio.gabbrielli@unibo.it).

Digital Object Identifier 10.1109/TLT.2024.3425959

We considered data on approximately 40 000 students. They cover multiple information, including demographics, prior schooling, enrollment, and first-year academic performance, to identify patterns in students' academic trajectories and predict dropout risk at an early stage. To facilitate early intervention, we include data on the academic performance of the same cohorts of students at different time intervals, i.e., at enrollment, after three, six, nine, and twelve months. By measuring the performance of the model at each time interval, we aim to answer the second RQ.

RQ2: To what extent does postenrollment academic career information improve model performance?

Given the current model of university assessment and education, we hypothesize that students' academic career characteristics can provide valuable insights into predicting dropout risk. The validity of this hypothesis is tested using our set of predictive models, trained at different stages, as a simulation tool. We use appropriate performance metrics, including precision, recall, and F1, to test and analyze the hypothesis. While predictive accuracy is commonly used to select data-driven solutions, we also recognize the importance of explainability in enhancing model trustworthiness and promoting its adoption [6], [7]. Explainability refers to the model's ability to provide transparent justifications for predictions, enabling stakeholders to understand the underlying factors [8]. By assessing both predictive accuracy and explainability, we ensure that outcome prediction models provide meaningful insights for informed decisions and interventions. To achieve reliable measures of importance, we compare different post hoc explainability techniques.

RQ3: To what extent do the explanations obtained from various post hoc explanatory techniques contribute to the reliability of the hypotheses underlying our models and their results?

Our experiments were conducted using all available data, including all students enrolled in any undergraduate course at the university over three academic years, with no subsampling or data exclusion.

The rest of this article is organized as follows. Section II provides an overview of related approaches. In Section III, we describe the dataset, introduce the predictive models, and provide an overview of the explanatory techniques used. The metrics of predictive performance are presented. Sections IV and V present the results, comparing predictive performance, and explanatory power under different assumptions. Section VI discusses the results in relation to the RQs. Finally, Section VII concludes this article.

II. Related Work

The study of understanding and decreasing dropout rates within higher education has advanced significantly, with numerous investigations utilizing diverse analytical methodologies and data sources [2], [9]. This review focuses on research in conventional *in-person* classrooms, categorized by ML algorithms (RQ1), the role of academic career information (RQ2), and the impact of post hoc explainability techniques (RQ3).

A. ML in School Dropout Prevention

The landscape of ML algorithms for academic dropout prediction has evolved significantly [10], [11], [12], with a growing emphasis on the adaptability and performance of deep architectures [13]. Early work by Anand et al. [14] used recursive clustering to evaluate student performance in programming courses, identifying underperforming students early. Alban and Mauricio[15] introduced neural networks for university dropout prediction, using multilayer perceptrons and radial basis function networks to achieve high accuracy rates. Nabil et al. [13] compared various ML algorithms, finding that DNNs outperformed traditional methods due to their ability to capture nonlinear correlations between student characteristics.

Baranyi et al. [16] extended the utility of deep learning by focusing on interpretability. They used DNNs and gradient-boosted trees, achieving high prediction accuracy and providing feature ranking through permutation importance and shapley additive explanations (SHAP) values. Tang et al. [17] introduced knowledge interaction discovery network (KIDNet), a knowledge-aware neural network model that combines factorization machine and DNN algorithms to capture both lower order and higher order feature interactions, demonstrating its effectiveness on a real-world dataset .

In summary, the empirical validation of deep architectures for predicting academic dropout has enriched the state of the art and opened avenues for future research. Our work aligns with this trend by adopting the FTT model [4], exploring the potential of attention-based neural networks for tabular data in the context of academic dropout prediction. This research aims to leverage these architectures to develop more effective and nuanced models to mitigate dropout rates.

B. Data Sources and Features for Predicting Academic Risk

We review the types of data sources and features used in existing literature to predict academic risk, focusing on academic history information. Dekker et al. [18] used structured data, such as student grades and attendance records, to predict dropout in electrical engineering programs, achieving 75% to 80% accuracy with decision trees. Kiss et al. [19] incorporated both structured and unstructured data, including preenrollment achievement measures and first-semester performance indicators, using artificial neural networks and boosting algorithms to highlight the incremental predictive validity of early university performance indicators.

Jayaraman [20] used unstructured data from counselor notes, employing natural language processing techniques to extract sentiments and using them as features in an RF model, achieving 73% accuracy in predicting student dropout. Del Bonifro et al. [10] presented a prediction tool that uses ML techniques to assess the risk of first-year undergraduate students, incorporating a range of variables from personal data to proficiency credits. This study serves as a foundational reference, particularly in its methodological approach to using preenrollment and first-year academic data for predictive modeling.

Alwarthan et al. [21] conducted a systematic review of data mining techniques used to predict student academic performance, identifying RF and ensemble models as the most accurate but noting a lack of consensus on the impact of admissions requirements on student performance. Alam [22] introduced a multimodal neural fusion network combining structured and unstructured data to predict various student retention risks, reporting promising performance and investigating the fairness of the model.

Our research aligns with the existing literature on using structured data for predictive modeling, capturing the temporal aspects of academic performance through a time-series approach. Our dataset comprises over 40 000 student careers, spanning three academic cohorts and including 110 different degree programs, enhancing the predictive power and generalizability of our models across different academic contexts.

C. Model Interpretability and Explainability in Education

The importance of interpretability and explainability in ML models is particularly pronounced in educational data mining, where the implications extend to human futures and career trajectories. Cohausz [23] emphasized the need for a nuanced, multistage approach to interpretability, advocating a fusion of AI and social science methodologies, and extending local interpretable model-agnostic explanations (LIME) [24] for deeper interpretation.

Cannistrà et al. [25] highlighted the pivotal role of feature relevance in early dropout prediction, using an information-driven modeling strategy and considering the specific programs in which students were enrolled. Nagy and Molontay [26] used a range of explainable AI (XAI) tools, such as permutation importance, partial dependence plots, LIME, and SHAP scores, demonstrating their utility in elucidating both global and local aspects of dropout prediction models. Delen et al. [27] presented a hybrid ML framework designed to provide actionable insights for individualized interventions, cautioning against the indiscriminate application of group-level insights for individual decision making.

In line with these contributions, our research highlights the criticality of model interpretability and explainability. As detailed in Section III-D, our methodology incorporates both global and local perspectives on explainability, emphasizing reliability and validity, underpinned by our comprehensive dataset and rigorous evaluation metrics.

III. Materials and Methods

A. Dataset Description

The dataset used for this work was extracted from a collection of real data from one of the largest Italian universities. Specifically, we have considered pseudonymous data describing 44 875 students enrolled in 110 courses in the academic years 2018–2019, 2019–2020, and 2020–2021. The dataset is collected by the university, thanks to the informed consent provided by students at the time of enrollment. This allows the data to be used in pseudonymized form for research activities aimed at improving the teaching offer and academic services. However, the pseudonymization of the dataset ensures that students cannot be identified, thereby meeting the ethical requirements of the research.

TABLE I
Available Features for Each Student in the Original Dataset, Along With the Possible Values Range

UId	Features	Type	Range
AE	Age of enrollment	Numeric	≥ 0
SG	Student gender	Nominal	1, 2
GOma	Geographical origin (macro)	Nominal	1–6
GOmi	Geographical origin (micro)	Nominal	1–76
EFSI	EFSI	Nominal	1–8
HST	High school type	Nominal	1–10
HSM	High school (final) mark	Numeric	60–100
CD	First/Single cycle degree	Nominal	1, 2
AS	Academic school ID	Nominal	1–11
DN	Degree name	Nominal	1–97
PT	Place of teaching	Nominal	1–9
ALR	Additional learning Reqs.	Nominal	1, 2, 3
WMA	Weighted marks average	Numeric	0 or 18–30
NH	Number of honors	Numeric	≥ 0
ECTS	Number of credits	Numeric	0–60
DO	Dropout	Nominal	True or False

The first column uniquely identifies the corresponding feature.

Our analysis focuses on the first year. Statistical evidence from the source data suggests a concentration of dropouts in the first year of the course, with the phenomenon gradually decreasing in subsequent years. For the 2018 cohort of students, the only one for which we have data three years after enrollment, the dropout rate after one year is 14.8% of the total number of enrolled students, while those who leave by the third year is 23.4%. This means that 63.2% of the registered dropouts occurred in the first year, confirming the importance of acting within the first year to prevent dropouts.

Table I provides a comprehensive overview of the features of the dataset. The table is divided into four columns: the first column serves as a unique identifier for each feature, which will be referenced later in Section V; the second column names the feature; the third column specifies its type (either nominal or numeric); and the fourth column outlines the possible values or ranges.

The features are categorized into four distinct groups.

1) *Personal data* includes characteristics such as gender, age, and geographical origin, as well as the equivalent economic situation indicator (EESI), which measures the economic status of the family at the time of enrollment.
   a) *Age of enrollment:* This numeric feature indicates the age of students at the time of their enrollment. It can offer insights into the relationship between age and academic performance or dropout rates.
   b) *Student gender:* This feature captures the gender classes given as binary (male or female) encoding. This is used as a basis for stratified analyses to assess model fairness across gender categories.
   c) *Geographical origin:* This feature is further divided into macro- and microcategorizations. The macrocategorization identifies six modalities, distinguishing between four macroareas in Italy, foreign students, and instances where the information is not available. The microcategorization offers 76 possible values, corresponding to either the Italian region or the country of origin for foreign students.

   d) *EESI:* The EESI feature is optional upon enrollment and is segmented into eight distinct financial bands. These bands are designed to encapsulate the economic status of both the students and their families. The bands are ordinal in nature, ranging from the lowest, which signifies the most financially disadvantaged situations, to the highest, indicative of more financially favorable conditions.

2) *Educational background* relates to the educational background attained at the upper secondary level. Specifically, this group includes the following two key characteristics.
   a) *High school type:* This nominal characteristic delineates ten different types of high schools from which students graduated. It is used to capture the diversity of educational backgrounds and to potentially elucidate any correlations between the type of high school attended and academic performance or dropout rates in higher education.
   b) *High school final mark:* This numerical characteristic represents the final mark obtained by students at the end of their high school education. It is intended to provide an initial quantitative measure of academic competence that may be indicative of subsequent performance in higher education.
3) *Academic program* set relates to the characteristics of the program in which the student is enrolled. This group comprises several attributes.
   a) *First/single cycle degree:* This ordinal characteristic categorizes the length of the program. A value of "1" represents first cycle degrees, which typically last three years, while "2" corresponds to single cycle degrees, which last five or six years.
   b) *Academic school ID:* This nominal feature identifies the academic school selected by the student. The dataset currently includes 11 different academic schools, each potentially offering a unique set of degree programs.
   c) *Degree name:* This characteristic serves as a unique identifier for the specific program chosen by the student, allowing for granular analysis of academic pathways.
   d) *Place of teaching:* This nominal characteristic indicates the geographical location of the program's headquarters, with nine different cities represented in the dataset.
   e) *Additional learning requirements (ALR):* This ordinal feature accounts for the possibility of mandatory additional coursework during the first academic year. Certain programs require an admission test, and failure to pass this test necessitates additional coursework and subsequent examinations. The ALR characteristic is coded as follows: "1" indicates programs without ALR; "2" indicates that the ALR exam was passed; and "3" indicates that the required ALR exam was not passed.
4) *Academic performance* set relates to measures that capture students' academic progress after enrollment. This group is informed by the following three main variables, each available at different time intervals.
   a) *Weighted marks average (WMA):* This numerical characteristic represents the average examination mark, weighted by the corresponding European credit transfer and accumulation system (ECTS) credits for each examination. In the context of the Italian academic evaluation system, exam marks range between 18 and 30. Consequently, the weighted average also falls within this interval. If a student has not passed any exams, this average is set to 0.
   b) *Number of honors:* This numerical characteristic quantifies the cases where an exam was passed with honors. Note that although honors are recorded, they do not affect the weighted average of exam grades.
   c) *Number of credits:* This numerical characteristic indicates the total number of ECTS credits earned by the student. The maximum number of ECTS credits that can be accumulated in a single academic year is 60.

The *target variable* for our predictive models is the "dropout" characteristic, represented as a Boolean variable with values of 0 and 1, encoding `False` and `True`, respectively. Specifically, a value of 0 is assigned to students who exhibit canonical academic outcomes, characterized by the continuation of their studies and the successful acquisition of course credits. Conversely, a value of 1 encapsulates three distinct noncanonical outcomes, each of which indicates a form of academic withdrawal. The first category includes students who formally abandoned their studies without transferring to other Italian programs. The second category includes students who have transferred to other programs within the same academic institution. The third category consists of students who left their current program to enroll in another university.

It is appropriate to categorize these three noncanonical outcomes as forms of dropout, as they all represent a deviation from the student's original academic trajectory. The differences between them lie solely in the subsequent choices that students make after dropping out. Moreover, these noncanonical outcomes collectively constitute a minority in the dataset, accounting for 23.4%.

In order to facilitate a nuanced analysis of students' academic progress, we have divided the data into five different time intervals, each of which captures a different phase of the first academic year. These intervals are defined at 0, 3, 6, 9, and 12 months after enrollment. Importantly, each student is represented in each of these intervals; no data points were excluded at any time. This approach resulted in five different versions of the dataset for each cohort of students. The versions are distinguished by the values of the fourth set of characteristics, which are updated to reflect the academic metrics at each time interval. This methodology allows us to strike a balance between making early predictions and capturing the evolving academic trajectories of students.

B. Predictive Models Explained

In our study, we employ two cutting-edge ML algorithms, each with distinct characteristics and advantages: RF and FTT.

These algorithms are chosen for their ability to handle complex, high-dimensional data, comprising various temporal snapshots reflecting students' academic progression.

RF [28] is an ensemble learning method renowned for its robustness and accuracy. At its core, RF creates a "forest" of decision trees, each trained on a random subset of the dataset. This technique, known as bootstrap aggregation or bagging, enhances the model's generalizability, effectively reducing the risk of overfitting. Overfitting occurs when a model learns the training data too well, including its noise, which can negatively impact its performance on unseen data. By aggregating predictions from multiple trees, RF produces a more stable and accurate prediction. To optimize our RF model, we employed a technique called grid search to meticulously adjust the model's parameters, ensuring the best possible performance. Model efficacy was gauged using fivefold cross validation, a process where the dataset is partitioned into five parts, allowing the model to be trained and tested across multiple scenarios to ensure reliability. Our evaluation metrics include balanced accuracy, which adjusts for any imbalance in the dataset's classes, and sensitivity, indicating the model's ability to correctly identify positive instances.

The FTT model [4] represents a novel application of neural network architecture in analyzing tabular data. Drawing inspiration from the transformer model [5], which has revolutionized natural language processing, the FTT employs an attention mechanism. This mechanism enables the model to focus on the most informative features of the data, adapting dynamically to the importance of different inputs. The "feature tokenizer" component of FTT transforms the input data into tokens, akin to words in a sentence, allowing the transformer to process them effectively. This architecture is particularly adept at uncovering intricate patterns and relationships within the data, potentially offering superior predictive performance compared to more traditional methods.

To put the performance of our chosen models into perspective, we also implemented a basic model that simply predicts the most frequent class for all instances. This naive baseline serves as a point of reference to evaluate the added value of the more sophisticated RF and FTT models, especially in the context of our dataset's class imbalance.

C. Dataset Preprocessing

The dataset contains both numerical and categorical characteristics. Numerical features, such as age at enrollment and final school grade, are processed as floating point numbers. The target variable for classification, called the "dropout" feature, is Boolean, as explained in the previous subsection.

Categorical features require different preprocessing techniques to adapt to the specificities of RF and FTT algorithms. For RF, one-hot encoding is used to fit the training set. Unknown categories are handled by a zero vector representation.

In contrast, the FTT models use label encoding, which assigns unique numerical labels to each category within a feature. This method is advantageous for algorithms that benefit from ordinal relationships between categories. Similar to RF, FTT models are trained on the training set, and the same encoding scheme is used for validation and testing. Unknown categories are coded as zero.

The dataset exhibits class imbalance, with dropout instances representing only 15.4% of the total. Such imbalance can negatively affect the performance of binary classification models [29].

For RF models, the imbalance is mitigated by class-weighted options. The weights are calculated based on the bootstrap sample for each decision tree and are inversely proportional to the class frequencies. These weights influence both the entropy criterion for splits and the "weighted majority vote" of the terminal nodes [30].

In the case of FTT, random weighted batch sampling is used to counteract the imbalance. This technique adjusts the selection probabilities based on class frequencies, thereby improving the representation of the minority class during training. This eliminates the need for data replication and mitigates the effects of class imbalance.

D. XAI Techniques

One of the main contributions of this work is the implementation of explainability techniques to understand the predictions made by RF and FTT models. We focus on computing feature importance, determining how each feature contributes to the predictions. We applied our explainability strategies to the top-performing model in each family: grouped permutation importance (GPI) for RF [31], attention map (AM) for FTT [4], [32], and SHAP [33] for both.

RF models are valued for their interpretability. We used GPI [34], an adaptation of permutation feature importance (PFI) [35], which addresses the consistency issue of one-hot encoded features by treating them as a single block during shuffling. GPI also incorporates feature weighting within each group to account for varying feature importance. This model-agnostic, post hoc technique can be applied universally to explain trained black-box models [8].

For FTT models, we used the AM to compute feature importance. This model-specific technique relies on the attention mechanism in transformers, averaging the attention weights for each token in the sample to determine feature importance.

In addition to GPI and AM, we employed SHAP [36], inspired by Shapley values from the cooperative game theory, to provide local explanations for individual predictions. SHAP quantifies the influence of each feature on a model's prediction, offering nuanced insights into feature contributions for each instance.

GPI and AM offer global explanations, identifying which features drive overall model performance. GPI measures feature importance by the decrease in the model's sensitivity, while AM uses attention weights to estimate feature usage by the model. SHAP provides local explanations, detailing the impact of features on individual predictions.

We used GPI and AM to derive global explanations from the test set for RF and FTT models, respectively, and compared the features identified by both techniques. SHAP was used for local explanations on three selected students (early dropout, transfer, nondropout) and for a global perspective using a beeswarm plot, summarizing how top features impact the model's output.

Due to the high computational cost of SHAP, we applied approximation strategies. Kernel SHAP [36] was used for general models, while Tree SHAP [33], optimized for decision trees, was applied to RF models. This allowed us to include all test instances in the RF beeswarm plot and limit FTT samples to 200 randomly selected instances.

E. Performance Metrics

To evaluate our prediction models, we employ several common performance metrics for binary classifiers: accuracy, sensitivity (recall), Sspecificity, and weighted F1 score. These metrics provide a comprehensive assessment, especially crucial in our imbalanced dataset.

Accuracy quantifies the proportion of correct predictions across both classes relative to the entire dataset. While useful, its reliability can be compromised in the presence of class imbalance, where the model may favor the majority class. Therefore, we use additional metrics to provide a more nuanced evaluation.

Sensitivity (or Recall) measures the model's ability to correctly identify students at high risk of dropping out. This metric is crucial because the high-risk group, although a minority, is of significant interest to educational stakeholders. High sensitivity ensures effective identification of students who need targeted interventions.

Specificity assesses the model's performance in correctly identifying students at low risk of dropping out. It complements sensitivity by indicating how well the model avoids false positives.

Weighted F1 score balances sensitivity (recall) and precision (positive predictive value), providing a harmonic mean of these two metrics. This measure is particularly useful in imbalanced datasets as it accounts for both false positives and false negatives, giving a comprehensive view of the model's performance. A high F1 score indicates that the model is both accurate and sensitive, effectively identifying students across risk categories.

F. Model Training and Validation

The dataset was split into training, validation, and test sets. We used data from the academic years 2018–2019 and 2019–2020 for training and validation, while data from 2020–2021 was reserved for testing. The split ratio for the training and validation sets was 70:30.

For training the RF models, we performed grid search cross validation to identify the optimal hyperparameters, including the number of trees, maximum depth, and minimum samples per leaf. The models were trained using fivefold cross validation to ensure robustness and generalizability, i.e., the dataset was divided into five subsets, the model was trained on four subsets and validated on the remaining one.

The FTT models were trained using the Adam optimizer with a learning rate schedule that decayed the learning rate based on the validation loss. We used early stopping to prevent overfitting, monitoring the validation loss and halting training when no improvement was observed for a set number of epochs. Random weighted batch sampling was employed to handle class imbalance during training.

To ensure the reproducibility of our results, we set random seeds for all random processes involved in data splitting, model training, and evaluation. All experiments were conducted using Python with libraries such as scikit-learn for the RF models and PyTorch for the FTT models.

G. Fairness Analysis

In addition to evaluating predictive performance, our study places a strong emphasis on the ethical integrity of the predictive models, particularly in terms of fairness [37]. Fairness analysis was conducted to identify and mitigate potential biases that may differentially impact specific demographic groups.

We focused on several protected attributes, including gender, geographical origin (categorized into macro regions), and economic status as indicated by the EESI. These attributes were chosen due to their relevance in reflecting the diverse backgrounds of the student population and their potential influence on academic outcomes.

For each protected attribute, we performed stratified analyses to evaluate the performance metrics across different subgroups. The metrics analyzed included accuracy, precision, sensitivity, specificity, false positive rates, and false negative rates. This analysis was aimed at detecting any disparities in model performance that could indicate bias.

To quantify fairness, we utilized several fairness metrics following recent pivotal research [37], [38] as follows.

1) *Demographic parity:* Ensures that the prediction rate is similar across different demographic groups.
2) *Equalized odds:* Requires that the true positive rate and false positive rate are similar across different demographic groups.
3) *Predictive parity:* Ensures that the precision is similar across different demographic groups.

Confidence intervals at the 95% level were determined using bootstrap resampling techniques. These intervals provide a measure of the variability in our fairness metrics and help in assessing the statistical significance of any observed disparities.

In the stratified analysis, we used the following steps:

1) define groups based on the protected attributes (e.g., male versus female for gender);
2) calculates the performance and fairness metrics for each group;
3) compare the metrics across groups to identify disparities and analyze the potential causes of any detected bias.

IV. Predictive Performance Results

In this section, we present the performance metrics, including accuracy and sensitivity, for both the RF and FTT models at different time intervals after enrollment, as described in Section III. These metrics are evaluated on the test set. The FTT models generally demonstrate superior accuracy compared to their RF counterparts, except when assessed six-months postenrollment. Conversely, the RF models show improved sensitivity capabilities under certain conditions.

TABLE II
SUMMARY OF ACCURACY AND SENSITIVITY ON TEST SET

	Accuracy		Sensitivity	
Time step	RF	FTT	RF	FTT
October enrolment (T0)	0.72	0.78	0.48	0.44
End of January (T1)	0.75	0.78	0.65	0.51
End of April (T2)	0.84	0.83	0.59	0.65
End of July (T3)	0.85	0.86	0.75	0.74
End of October (T4)	0.85	0.87	0.80	0.81

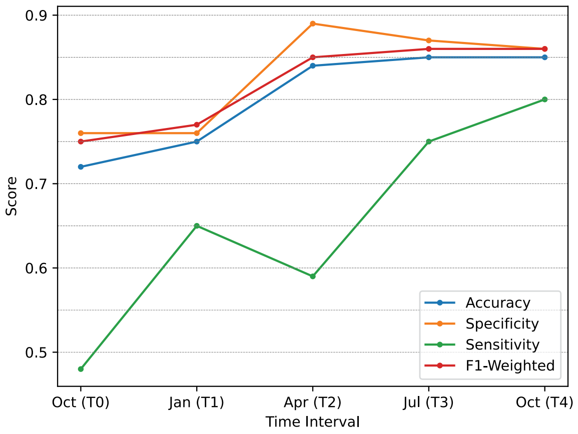

Fig. 1. *RF model performance over time*. Variation in accuracy, sensitivity, specificity, and weighted F1-score at different intervals from October enrollment. Actual weighted F1 score values are provided.

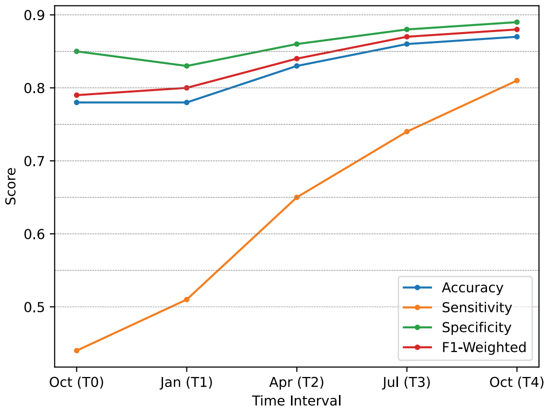

Fig. 2. *FTT model performance over time*. Fluctuations in accuracy, sensitivity, specificity, and weighted F1-score at specific intervals from October enrollment. Actual weighted F1 score values are included.

As shown in Table II, the best-performing model overall is the FTT variant trained on data available 12 months after enrollment, achieving an accuracy of 0.87 and a sensitivity of 0.81.

For comparative analysis, we also introduce a naive baseline classifier that predicts the majority class label from the training set across all test instances. This classifier achieves an accuracy of 0.84 and a sensitivity of 0.25 over all time intervals considered.

Figs. 1 and 2 illustrate the time trends in the performance metrics for the RF and FTT models. The weighted F1 score, identified as the most equitable metric in Section III-E, also shows a general improvement over time. This suggests that the quarterly updates of student career information contribute significantly to the predictive power of the models.

For RF models, the most notable improvement in performance occurs between the zero- and six-month intervals, with a slight decrease in sensitivity thereafter. We briefly discuss the behavior of sensitivity trend in Section V. For the FTT models, the metrics show a more consistent upward trend, reaching satisfactory levels even at the time of enrollment.

The performance of the naive classifier serves as a baseline to help interpret the effectiveness of the RF and FTT models, particularly in scenarios with unbalanced datasets.

A. *Fairness Analysis Results*

In Fig. 3, we present a comprehensive comparison of the fairness analysis of the best model in terms of predictive performance. The figure consists of 15 box plots, systematically arranged in a grid. Each row in this grid is dedicated to the analysis of a particular feature, while each column corresponds to one of the selected evaluation metrics—namely accuracy, recall, precision, false positive rate, and false negative rate—all evaluated at a decision threshold of 0.5. This visual representation serves as a robust tool for examining the performance of the model across different subgroups, thereby facilitating a nuanced understanding of its fairness attributes.

We performed the fairness analysis by segmenting the dataset based on specific demographic and academic features. The segmentation allowed us to evaluate how the model performs across various subgroups, ensuring that no particular group is disproportionately advantaged or disadvantaged. The fairness analysis ensures that our predictive models not only achieve high accuracy but also uphold ethical standards essential in educational settings.

V. EXPLAINABILITY RESULTS

In this section, we outline the results of our explainability analysis, first adopting a global XAI framework similar to the methodology presented in [35], and then, extending our investigation through the application of localized techniques. Among the RF-based models, we choose two versions: the model trained with data six months after enrollment (referred to as T2-RF model hereinafter) and the one trained with data 12 months after enrollment (T4-RF model). The first analysis aims to get insights into why the model registered a pitfall in the sensitivity performance to the advantage of specificity (see Fig. 1). The second model has been chosen because it has the highest results according to all the performance metrics. For FTT, we considered the 12-months model (T4-FTT model), which is our best model according to the results presented in Section IV.

Our explainability results are organized as follows. In Section V-A, we present the global explainability perspective with the techniques chosen for each model, i.e., GPI and AM for RF and FTT, respectively. In Section V-B, we present the results

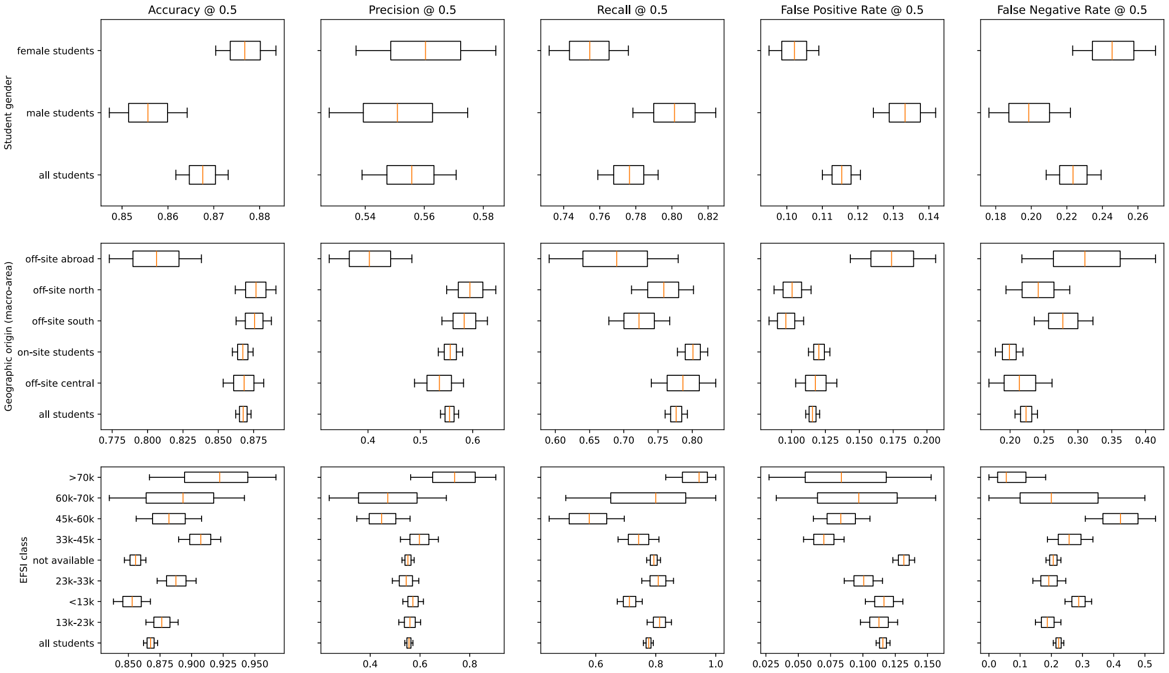

Fig. 3. *Quantitative analysis of fairness across multiple features and metrics*. The boxplot matrix is organized into three rows (student gender, geographical origin, and economic class) and five columns (accuracy, recall, precision, false positive rate, and false negative rate), all evaluated at a decision threshold of 0.5. This layout provides a comprehensive view of model fairness across different subgroups and evaluation criteria.

TABLE III
GPI Results for RF Computed as Mean Decrease of Sensitivity

Feature	T2-RF	T4-RF
Weighted mark average	0.248 (0.007)	0.001 (< 0.001)
Number of ECTS	0.036 (0.003)	0.493 (0.005)
Additional learning reqs.	0.025 (0.003)	0.005 (0.002)
Academic School	0.019 (0.002)	0.005 (0.001)
Age of enrollment	0.001 (< 0.001)	0.005 (0.003)

obtained with SHAP when used in its local explainability mode. We present its application to students in different conditions of continuation of studies as an example of the kind of insights that can be derived locally with SHAP. Finally, in Section V-C, we present beeswarm plots for a global perspective through SHAP values, both for RF and FTT.

A. Global Feature Importance

The use of GPI in the RF model facilitates the identification of salient features that contribute to the generation of predictions for each trained instantiation of the model. In this study, a feature is considered significant if its importance measure is greater than or equal to 0.01. This corresponds to a minimum 1% decrease in sensitivity due to random shuffling of that particular feature. Conversely, a feature is considered negligible if its importance measure falls below the 0.01 threshold. We performed 100 random shuffles for each feature to calculate the PFI. The mean and standard deviation of the most salient features for the T2-RF and T4-RF models are shown in Table III.

In the T2-RF model, the most salient feature is the weighted mean grade, denoted by $\text{mean}_{\text{WMA}} = 0.248$, followed by the number of ECTS credits earned within six months of enrollment, denoted by $\text{mean}_{\text{ECTS}} = 0.036$; i.e., they contribute to a decrease in sensitivity of 25% and 4%, respectively. The threshold of 1% is also exceeded by the allocation of ALR, which are determined on the basis of admission tests, and by the categorization of the academic school.

For the T4-RF model, we visualized an equivalent number of features as identified for the T2-RF model. However, only the number of ECTS credits earned 12 months after enrollment exceeds the 1% threshold criterion. Specifically, this feature shows an average decrease of 50% in the sensitivity metric, and thus, emerges as the most important variable for identifying dropout risk. For the remaining features, the perturbation in sensitivity due to randomized reshuffling is insignificant, falling below the 1% threshold criterion.

The key difference between the two models is the impact of WMA and ECTS on sensitivity. As ECTS range increases along the academic year and WMA range remains stable, ECTS's importance grows relative to WMA. The results (see Fig. 1) indicate that reliable ECTS information, available by the end of the first enrollment year, is crucial for model robustness compared to WMA. Indeed, where WMA matters the most, according to GPI analysis, there is a pitfall in sensitivity.

For the FTT model, we applied an AM-based feature importance analysis to the T4-FTT model. In Fig. 4, the average weight assigned to each feature in the AM is shown. It is noteworthy that no single feature has a significantly higher average weight than the others. The average weights for all features are in the range $[0.0494, 0.0626]$. Nevertheless, it is worth noting that the top five features, in descending order of importance, are the geographical region of origin, the place of teaching, the number

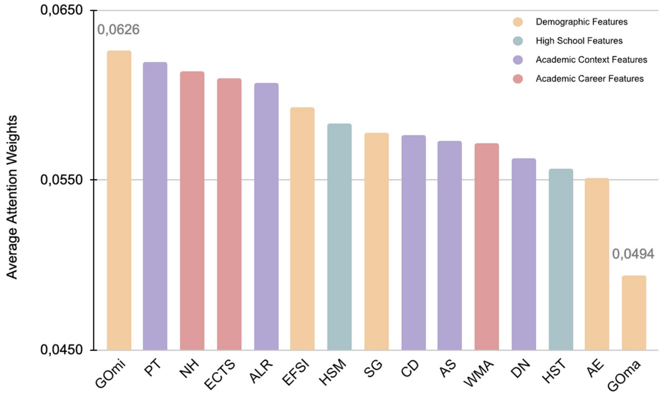

Fig. 4. *AM-based feature importance for T4-FTT model*. Each bar shows the average attention weight for a feature in the training dataset, colored by information type as introduced in Section III.

of awards obtained, the number of ECTS credits obtained, and information on ALR.

The two global explainability techniques applied to their respective models (T4-RF and T4-FTT) provide different insights. ECTS is by far the preeminent feature for T4-RF with GPI; on the other hand, it ranks fourth for T4-FTT with AM feature importance in a context where no feature stands out more than the others. However, the relevance of features related to the student's current academic career is consistent across both models. While the global feature importance procedure offers a preliminary understanding of feature importance, it is worth noting that the results may not fully capture the explainability power of the models. The limitations of this kind of analysis suggest that more sophisticated techniques, such as SHAP, could be employed to provide a more comprehensive and interpretable understanding of the models' decision-making processes.

B. *SHAP for Local Explanations*

The SHAP explainability technique has been applied to both RF and FTT models. We chose SHAP because it is an agnostic state-of-the-art explainability technique. Thus, we considered the best models both for RF and FTT, i.e., T4-RF and T4-FTT models, and we present and compare their explainability outcomes. First, we aim to introduce the results of local explainability gained from the models on some selected students, taken as examples. Figs. 5 and 6 display how the selected models came to the prediction correctly for three selected students, i.e., the predicted risk for the presented cases agrees with their actual value. Fig. 5 refers to the RF model, and Fig. 6 refers to the FTT one. In each figure, we selected a student who early interrupts the academic career, a student who transfers to another degree program, and a student for whom dropout does not occur.

As for local explainability with the RF model, the number of ECTS is the one with the highest SHAP value (longest bar in the plot) both for the student who early interrupts the academic career (case *a* in Fig. 5) and for the one for whom dropout did not occur (case *c*). The bar color and its orientation tell how this feature contributes to the predicted risk: for student *a*, pink and left-right oriented ECTS bar, not having accrued credits in 12 months raises the risk of dropout; for student *c*, blue and right-left oriented ECTS bar, having acquired 42 ECTS (out of 60 total) contributes to the prediction of a low risk of dropout. The same feature acts misleadingly for student *b*. In this case, ECTS is the second main feature according to its SHAP value. The attainment of 40 out of 60 ECTS credits is considered to be satisfactory as per the model. Typically, 60 ECTS credits represent the maximum amount of credits that a student can earn during the first year of enrollment. Thus, it is used to downgrade the dropout risk prediction, although the actual target class for the student is positive to dropout. The most relevant feature for high-risk dropout for the student *b* is the academic school, whose actual value is pharmacy and biotechnology. Statistics confirm that this academic school is affected by the highest number of transfers compared to other schools of the same university (36.9$\%$ in the three-year enrollment period 2018–2021 against a university average of 8.4$\%$). This is because many first-year students choose pharmacy and biotechnology courses as a second study choice after being excluded from other degrees with restricted admission procedures, e.g., medicine and surgery, or veterinary medicine. As a final remark for the RF model, also WMA appears as a relevant feature for the dropout predictions for all the students (among the first five SHAP values).

As regards local explainability with the FTT model, we refer to the examples in Fig. 6, and introduce the enabled explanations also in comparison with our observations for the RF model. Also for the FTT model, ECTS is the prominent feature in the risk prediction for students *a* and *c*. The same feature is less relevant for the student *b* (it appears as the seventh positive SHAP value). We have previously noted, based on the global analysis of feature importance using the AM method, that ECTS stands out as one of the most significant features, despite not being favored by the RF models. The different SHAP value of ECTS on different samples fits with this result. Furthermore, we want to underline that, unlike what was observed for the corresponding case for the RF model, the number of ECTS acquired by the student *b* here contributes to raising the dropout risk, despite being an acceptable asset (40 out of 60). For student *a*, together with ECTS, the features associated with higher SHAP values are the assignment of ALR that have not been passed and the WMA (equal to zero as no ECTS has been acquired). All these factors contribute, as expected, to raise the risk of dropout. The prominent feature for student *b* is DN, which identifies the degree program. We found matching information for the RF model (related to the academic school), and we have already motivated how to interpret these results with some descriptive statistics. For student *c*, ECTS is definitely the prominent feature, followed by data on the academic school, which is engineering and architecture.

To sum up, we find two main similarities between the local explanations gained by the two models. First, the relevance of ECTS for the dropout risk prediction of students *a* and *c*. Second, the relevance of information on the context of enrollment for the student *b*, i.e., the enrollment in pharmacy. On the other hand, we have a main difference in how the number of ECTS (40 out of 60) is used for the dropout risk prediction of the student *b*. One might wonder what interpretation to give to

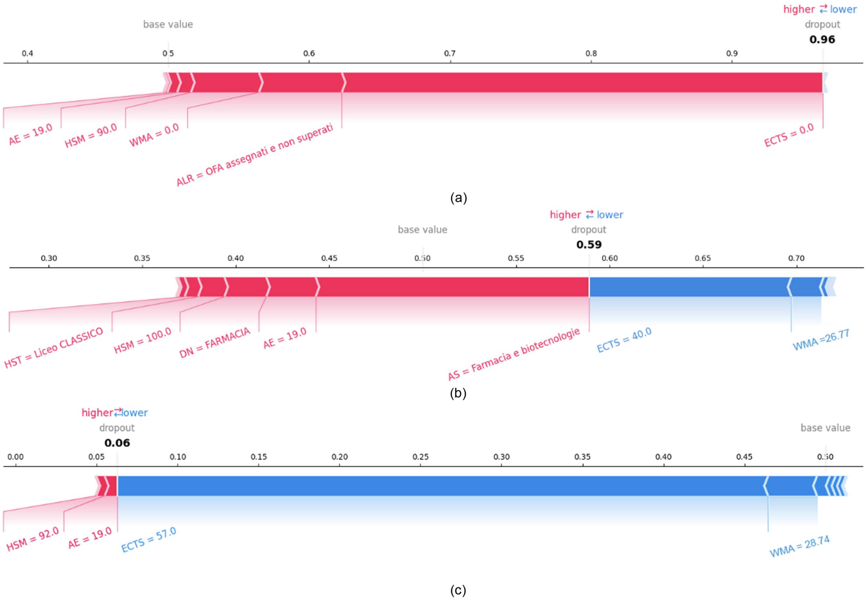

Fig. 5. *SHAP Local explanations for RF model trained with data twelve months after enrollment*. Each line shows the main features impacting the predicted dropout risk for a student, with bar lengths proportional to their SHAP values. Pink bars indicate features that increase dropout risk, while blue bars indicate features that decrease it. The combined contributions determine the predicted value. (a) Student who early interrupts the academic career. (b) Students who transfers to another degree course. (c) Student with no form of dropout.

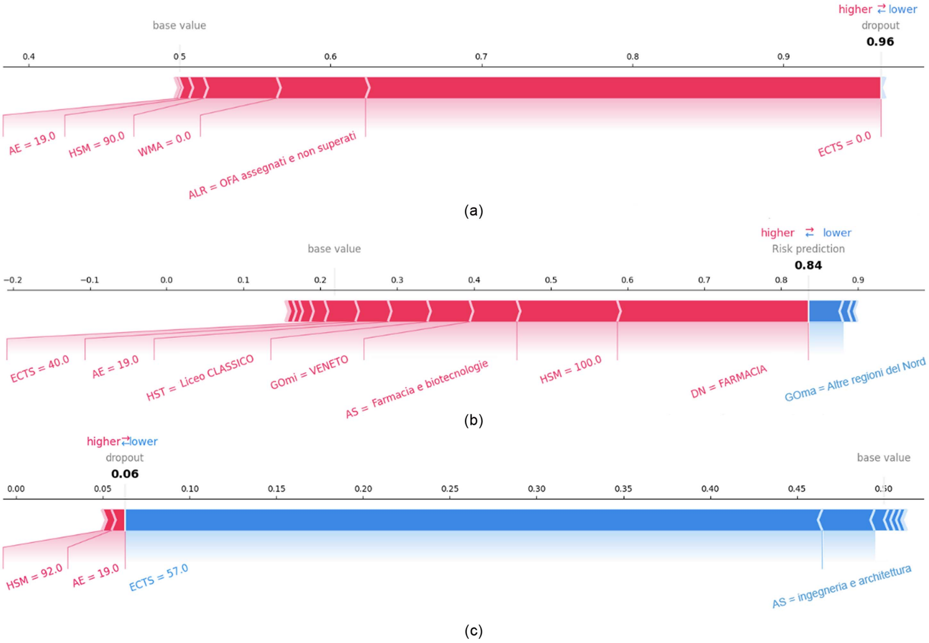

Fig. 6. *SHAP Local explanations for FTT model trained with data twelve months after enrollment*. Each line shows the main features impacting the predicted dropout risk for a student. Refer to Fig. 5 for instructions on reading the graph, which is similar for the RF model. (a) Student who early interrupts the academic career. (b) Students who transfers to another degree course. (c) Student with no form of dropout.

this difference. We hypothesize that the RF model struggles more in learning correlations between different features; the feature tokenizer module for input features embedding and the attention mechanisms of the FTT architecture provide greater flexibility, which allows, in the case of student *b*, to consider in a "contextualized" way the weight and the orientation effect of the number of ECTS. We deepen this discussion in Section VI.

C. *SHAP for Global Explanations*

Let us move back again to a global explainability perspective, aggregating local explanations computed with SHAP in a summary plot, namely beeswarm. Figs. 7 and 8 refer, respectively, to RF and FTT models, trained with data on students' academic careers 12 months after enrollment. In a beeswarm plot, for each instance, i.e., a student in our case study, the provided

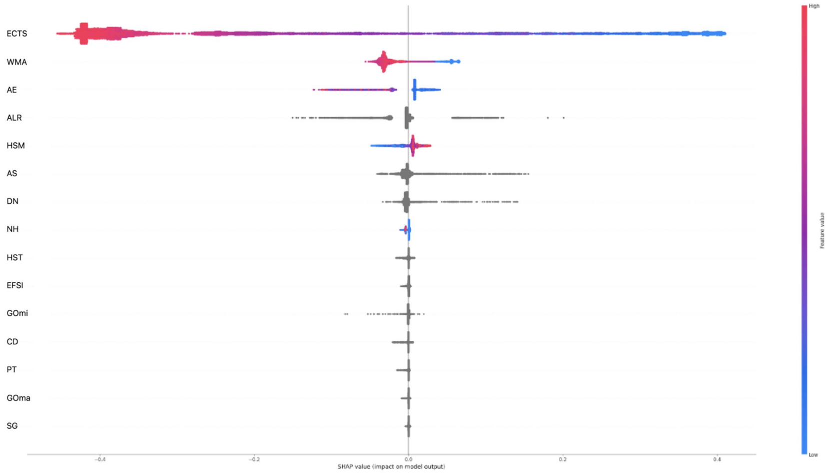

Fig. 7. *SHAP global explanations for RF model trained with data 12 months after enrollment*. The beeswarm plot for the T4-RF model shows features ordered by average SHAP value. Each dot represents an instance, positioned by SHAP value; colors indicate numeric feature values.

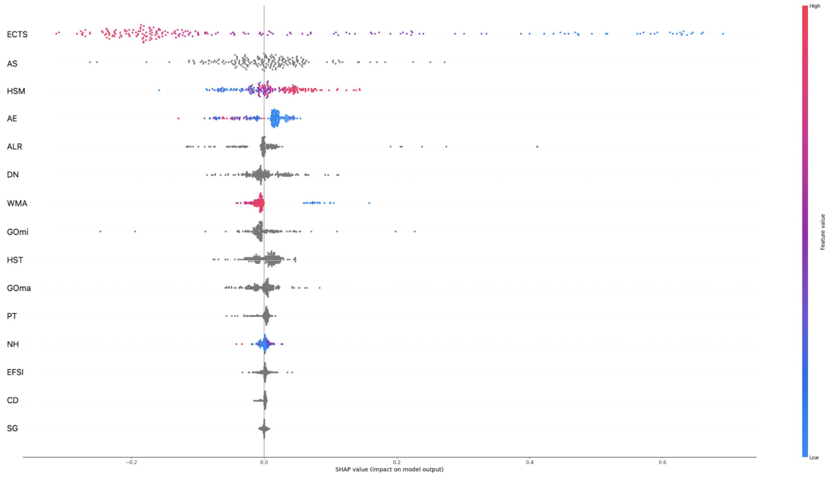

Fig. 8. *SHAP global explanations for FTT model trained with data 12 months after enrollment*. The beeswarm plot for the T4-FTT model shows features ordered by average SHAP value for 200 randomly chosen samples. Refer to Fig. 7 for guidance on interpreting the chart.

explanation is visualized by a single dot on each feature row. The SHAP value of the row feature for each instance determines the horizontal position of the dots, whose distribution along each row shows a density graph. This information may be exploited to provide a global overview of the feature's importance. Features are in descending order according to their mean SHAP value. Moreover, each dot for numerical features is colored according to a chromatic scale to display the original value of a feature for each instance. Thus, a blue point on the right side of the ECTS row means that there is a student with a low number of acquired ECTS and this has a great influence in boosting her/his risk of dropout.

As for categorical features, both binary and nonbinary, we consider each of them as a single factor of analysis, encompassing all its modes. Specifically, SHAP values are computed for each instance by summing the SHAP values of all binary features associated with that categorical feature. For this set of features, no color mapping has been set to avoid implying an order among categorical features.

The beeswarm plot for the T4-RF model points out ECTS as the main feature; this is in line with the result retrieved with GPI. In particular, there is a high density of pink dots on the left, thus we can infer that the model often uses the acquisition of a high number of ECTS as an impact criterion to place the student in the low-risk class. The weighted average mark is another determining factor for the low-risk class, considering the high-density area of pink dots, i.e., high weighted average mark, on the left side. For the applicability of the model, it would be of interest to determine the features that are most decisive for the high-risk class, i.e., we are looking for high-density areas in the right part of the plot. However, no such situation is evident in any row. We may observe a slight correlation between low age of enrollment or low weighted average mark with high dropout risk. It is worth noting a counterintuitive result among the explanations used by the model: the final high school grade (HSM) when high, is often considered by the RF model as a rationale for high dropout risk. Nevertheless, its impact on the prediction is small according to SHAP values. Finally, we point out the relevance that numeric features play in the RF model with respect to the categorical ones. The top three consists of all numeric features, and this is even more impressive if we consider that the dataset has only five numeric features against ten categorical ones. Among the categorical features, the most effective (fourth position according to mean SHAP value ranking) is the one on the attribution of ALR.

We conduct a similar analysis also for the T4-FTT model. We limited the beeswarm plot generation to 200 randomly chosen samples, due to time computational cost. This represents a limit in the SHAP global perspective on FTT but still allows us to obtain some insights. Similarly to the case of T4-RF, ECTS is the preeminent feature with a cluster of pink dots on the left side of the plot. Thus, in many cases, there is a correlation between a high achievement of ECTS and a lower dropout risk prediction. Other relevant data are those on the academic school of the students, their high school marks, age of enrollment, and information on ALR. We observe a blue cluster for the age of enrollment on the right side of the plot, revealing that young students are more likely to drop out. Also for the FTT model, the numerical features are relevant in determining the risk prediction. However, their distribution is less asymmetric than we observed for the RF model. In Fig. 7 for the T4-RF model, we have four out of the first five features that are numerical. Moreover, the numerical features are all in the first half of the ordered features. In the T4-FTT model, we have a more homogeneous, albeit still not symmetrical, distribution.

VI. Discussion

This section synthesizes the predictive and explanatory evidence to address the RQs outlined in the introduction, incorporating findings from similar studies and reflecting on the practical implications of our results.

In response to our first research question (*RQ1*), our empirical study supports the effectiveness of FTT models in predicting academic dropout risk. As detailed in Section IV, the FTT models, particularly the T4-FTT, consistently outperform RF models across various evaluation metrics, showing at least a one percentage point improvement. These results align with the trend of using deep learning algorithms for dropout prediction, recognized for their adaptability and sophistication [10], [11], [12], [13]. The balanced performance of FTT models in sensitivity and other metrics makes them suitable for dropout intervention strategies, consistent with findings in [14] and [15]. Furthermore, our fairness analysis, which shows consistent results across different characteristics within a 95% confidence interval, complements the focus on interpretability seen in [16].

Our study expands the use of deep learning in educational data mining, achieving high accuracy while improving model transparency and fairness. The flexible architecture of FTT models, including an embedding component and attention mechanism, allows customization to different feature sets and data distributions, enhancing predictive fairness. This novel application of attention-based neural networks for tabular data in dropout prediction adds to the existing literature [4], confirming the efficacy of FTT models for this task.

Regarding our second research question (*RQ2*), our findings highlight the crucial role of ECTS credits in predicting dropout risk, supporting prior research on the importance of academic history information [18], [19], [20]. Using structured data, including ECTS credits, our approach accurately predicts dropout risk, validated by our extensive dataset capturing diverse academic paths. The temporal sensitivity of our models to academic career characteristics underscores the dynamic nature of academic risk factors, with incremental improvements in predictive performance using data from progressively distant time points from enrollment. This aligns with Kiss et al.'s[19] emphasis on early university performance indicators.

However, relying solely on academic characteristics such as ECTS credits, which can change over time, has limitations. Building on the work presented in [20] and [22], we propose incorporating immediate behavioral features for a more comprehensive assessment of dropout risk. Our study verifies ECTS credits as a reliable dropout risk indicator and underscores the significance of analyzing academic career characteristics over time. Combining structured data in a time-series format, our methodology contributes to the current academic discourse, proposing ways to incorporate other information types for more comprehensive predictive models.

In examining research question 3 (*RQ3*), we explore the explainability of predictive models within educational data mining. Our findings align with and expand upon the current discourse on model interpretability . The significance of ECTS credits across different explainability techniques, such as global post hoc interpretability (GPI) and SHAP, supports previous studies [25], [26]. The detailed explanations provided by SHAP for the FTT model contribute to discussions about the adaptability of models to individual student cases, a topic also explored by Cohausz [23] and Delen et al. [27].

While both RF and FTT models focus heavily on ECTS credits, their ability to explain outcomes differs, highlighting the complexity of model interpretability. Our analysis showcases the FTT model's local explanatory power, especially regarding ECTS features, and emphasizes its versatility in adjusting to various feature sets and samples, allowing for a detailed understanding of factors contributing to dropout rates.

Our study also focuses on improving the sensitivity of the T2-RF model by evaluating the relationship between contextual information and academic career data. Incorporating ECTS credits as a feature enhances the understandability and reliability of predictive models, as highlighted by Delen et al. [27]. By transparently utilizing ECTS credits in our predictions, we improve the credibility and transparency of these models, emphasizing the importance of fairness and adaptability in their implementation.

VII. Conclusion

This study contributes to the evolving discourse on academic dropout prediction by employing two distinct ML methodologies: RF and FTT. Our investigation, driven by the dual aims of evaluating the comparative efficacy of FTT against conventional ML approaches such as RF and assessing the impact of incorporating academic career data, demonstrates the potential of ML to identify students at risk of dropout.

Our findings reveal that FTT models exhibit superior predictive accuracy over RF models, although with increased computational demands. Notably, the inclusion of academic career data enhances model performance, particularly in terms of sensitivity, and enriches the FTT models' capacity to profile students prone to dropout. Unlike RF models, which require distinct training for diverse outcomes, FTT models provide a more integrated approach to understanding dropout predictors.

The implications of our research for educational stakeholders are significant. By leveraging data-driven insights, institutions can better tailor their retention strategies. Specifically, integrating comprehensive datasets and employing advanced models such as FTT can provide a nuanced understanding of student behaviors and risk factors. This aligns with the findings of Zingaro et al. [6], who suggest that such approaches could refine retention strategies when combined with simulation-based analyses.

Despite these advances, our study highlights areas for further exploration. The reliance on ECTS credits, while informative, may obscure underlying causes of dropout. Future research should incorporate qualitative insights, such as student motivations and study habits, to provide a more holistic view of dropout causes. In addition, the development of tailored predictive models to address the unique dropout dynamics across different academic programs remains crucial.

Ethical considerations are paramount when utilizing predictive models in education. Future research should consider evaluating entire academic programs rather than individual students, incorporating broader contextual factors such as environmental and support mechanisms, as emphasized in [39]. By integrating insights from educational, cognitive, and psychological research into data-driven methodologies, we can enhance the ethical and effective application of ML in educational contexts, leading to a comprehensive understanding of academic dropout trends.

In summary, our study underscores the importance of incorporating diverse data sources and advanced ML models to improve dropout prediction. This approach not only enhances predictive accuracy but also ensures that interventions are informed, ethical, and effective, ultimately contributing to a deeper understanding of academic attrition and strategies to mitigate it.

Acknowledgment

In the interest of maintaining the highest standards of research integrity and transparency, the authors confirm that all experiments and analyses conducted in this study are fully reported in the article, with explicit disclosure of all measures, conditions, and controlled variables. No data were excluded from the analyses, and the study includes all first-year students enrolled during the data collection period, representing the entire population of interest.

References

[1] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Rev. Educ. Res.*, vol. 45, no. 1, pp. 89–125, 1975.

[2] T. Cardona, E. A. Cudney, R. Hoerl, and J. Snyder, "Data mining and machine learning retention models in higher education," *J. College Student Retention, Res. Theory Pract.*, vol. 25, 2020, Art. no. 1521025120964920.

[3] C. Beaulac and J. S. Rosenthal, "Predicting university students' academic success and major using random forests," *Res. Higher Educ.*, vol. 60, pp. 1048–1064, 2019.

[4] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18932–18943.

[5] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[6] S. Zingaro, A. Del Zozzo, F. Del Bonifro, and M. Gabbrielli, "Predictive models for effective policy making against university dropout," *Form@re, Open J. Per La Formazione Rete*, vol. 20, no. 3, pp. 165–175, Dec. 2020.

[7] D. Hooshyar and Y. Yang, "Neural-symbolic computing: A step toward interpretable AI in education," *Bull. Tech. Committee Learn. Technol.*, vol. 21, no. 4, pp. 2–6, 2021, ISSN: 2306-0212.

[8] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[9] R. Z. Pek, S. T. Özyer, T. Elhage, T. ÖZYER, and R. Alhajj, "The role of machine learning in identifying students at-risk and minimizing failure," *IEEE Access*, vol. 11, pp. 1224–1243, 2023.

[10] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, "Student dropout prediction," in *Artificial Intelligence in Education*, I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, and E. Millán, Eds. Cham, Switzerland: Springer, 2020, pp. 129–140.

[11] B. Prenkaj, P. Velardi, G. Stilo, D. Distante, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Comput. Surv.*, vol. 53, no. 3, May 2020, Art. no. 57.

[12] K. Fahd, S. Venkatraman, S. J. Miah, and K. Ahmed, "Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature," *Educ. Inf. Technol.*, vol. 27, no. 3, pp. 3743–3775, Apr. 2022.

[13] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021.

[14] V. Anand, S. A. Rahiman, E. B. George, and A. Huda, "Recursive clustering technique for students' performance evaluation in programming courses," in *Proc. Majan Int. Conf.*, 2018, pp. 1–5.

[15] M. Albán and D. Mauricio, "Neural networks to predict dropout at the universities," *Int. J. Mach. Learn. Comput.*, vol. 9, pp. 149–153, 2019.

[16] M. Baranyi, M. Nagy, and R. Molontay, "Interpretable deep learning for university dropout prediction," in *Proc. 21st Annu. Conf. Inf. Technol. Educ.*,, 2020, pp. 13–19.

[17] T. Tang, J. Hou, T. Guo, X. Bai, X. Tian, and A. N. Hoshyar, "KIDNet: A knowledge-aware neural network model for academic performance prediction," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2021, pp. 37–44.

[18] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proc. 2nd Int. Conf. Educ. Data Mining*, 2009, pp. 41–50.

[19] B. Kiss, M. Nagy, R. Molontay, and B. Csabay, "Predicting dropout using high school and first-semester academic achievement measures," in *Proc. 17th Int. Conf. Emerg. eLearning Technol. Appl.*, 2019, pp. 383–389.

[20] J. Jayaraman, "Predicting student dropout by mining advisor notes," in *Proc. 13th Int. Conf. Educ. Data Mining*, 2020, pp. 629–632.

[21] S. A. Alwarthan, N. Aslam, and I. U. Khan, "Predicting student academic performance at higher education using data mining: A systematic review," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, 2022, pp. 1–26.

[22] M. A. U. Alam, "College student retention risk analysis from educational database using multi-task multi-modal neural fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 12689–12697.

[23] L. Cohausz, "Towards real interpretability of student success prediction combining methods of XAI and social science," in *Proc. 15th Int. Educ. Data Mining Soc.*, Durham, U.K., Jul. 24–27, 2022.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144.

[25] M. Cannistrà, C. Masci, F. Ieva, T. Agasisti, and A. M. Paganoni, "Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques," *Stud. Higher Educ.*, vol. 47, no. 9, pp. 1935–1956, 2022.

[26] M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," *Int. J. Artif. Intell. Educ.*, vol. 34, pp. 274–300, 2023.

[27] D. Delen, B. Davazdahemami, and E. R. Dezfouli, "Predicting and mitigating freshmen student attrition: A local-explainable machine learning framework," *Inf. Syst. Front.*, vol. 26, pp. 641–662, 2023.

[28] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[29] Z. Zheng, Y. Li, and Y. Cai, "Oversampling method for imbalanced classification," *Comput. Inform.*, vol. 34, pp. 1017–1037, 2015.

[30] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse, "An empirical study of learning from imbalanced data using random forest," in *19th IEEE Int. Conf. Tools Artif. Intell.*, 2007, vol. 2, pp. 310–317, doi: 10.1109/ICTAI.2007.46.

[31] L. Plagwitz, A. Brenner, M. Fujarski, and J. Varghese, "Supporting AI-explainability by analyzing feature subsets in a machine learning model," in *Challenges of Trustable AI and Added-Value on Health*. Amsterdam, Netherlands: IOS Press, 2022, pp. 109–113.

[32] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič, "Feature importance estimation with self-attention networks," in *Proc. 24th Eur. Conf. Artif. Intell.*, Santiago de Compostela, Spain, 2020, vol. 325, pp. 1491–1498, doi: 10.3233/FAIA200256.

[33] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*.

[34] Q. Au, J. Herbinger, C. Stachl, B. Bischl, and G. Casalicchio, "Grouped feature importance and combined features effect plot," *Data Mining Knowl. Discov.*, vol. 36, no. 4, pp. 1401–1450, 2022.

[35] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, Apr. 2010.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[37] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness Accountability Transparency*, 2019, pp. 220–229.

[38] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 3323–3331.

[39] S. Matz, C. Bukow, H. Peters, C. Deacons, A. Dinu, and C. Stachl, "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," *Sci. Rep.*, vol. 13, 2023, Art. no. 5705.

Andrea Zanellati received the B.S. and M.S. degrees in mathematics from the University of Bologna, Bologna, Italy, in 2009 and 2011, respectively, where he is currently working toward the Ph.D. degree in data science and computation.

He has a permanent position as a Mathematics and Physics Teacher with the Italian High School System. His research interests include the application of data science and machine learning methods and techniques in education and large-scale assessment testing.

Stefano Pio Zingaro received the Ph.D. degree in computer science from the University of Bologna, Bologna, Italy, in 2020.

He has been an Assistant Professor of computer science with the Department of Computer Science and Engineering, University of Bologna, since 2023. His research interests include programming languages, distributed systems, and artificial intelligence.

Maurizio Gabbrielli received the Ph.D. degree in computer science from the University of Pisa, Pisa, Italy, in 1992.

He has been a Full Professor of computer science with the Department of Computer Science and Engineering, University of Bologna, Bologna, Italy, since 2001, where he has also been the Head of the Department, since 2021. He is also an Associate Dean for AI and Digital Soul and the Director of the master in management of digital technology with Bologna Business School, and a member of the INRIA FOCUS project team. He has worked at Centrum Wiskunde and Informatica, Amsterdam, Netherlands, the University of Pisa, and the University of Udine, Udine, Italy. His research interests include programming languages, constraint programming, and artificial intelligence.

Open Access funding provided by 'Alma Mater Studiorum – Università di Bologna' within the CRUI CARE Agreement