technologies

MDPI

Article

Evaluating and Forecasting Undergraduate Dropouts Using Machine Learning for Domestic and International Students

Songbo Wang * and Jiayi He

School of Civil Engineering, Architecture and Environment, Hubei University of Technology, Wuhan 430068, China
* Correspondence: wangsongbo@hbut.edu.cn

Abstract

Undergraduate dropout is a multidimensional phenomenon with implications for higher education, economic development, and social and cultural transformation, posing complex challenges for society as a whole. To address this, universities require effective dropout risk assessments for both domestic and international students, enabling the implementation of tailored strategies and support. This study sourced a dataset from multiple faculties, comprising 3544 records for domestic students (Portuguese) and 86 for international students, considering 23 features. To balance the data, Conditional Tabular Generative Adversarial Networks were utilized to generate 487 synthetic samples with comparable statistical characteristics for training (85%) while retaining the original 86 real samples for testing (15%), thus maintaining an identical train–test split for evaluating domestic students. An Automated Machine Learning framework, employing ensemble learning algorithms, achieved outstanding performance, with the Light Gradient Boosting Machine proving the most effective for domestic students and Categorical Boosting for international students, both achieving test accuracies exceeding 0.90. The analysis revealed that improving academic performance during the first and second semesters was key to reducing dropout risks. Once a satisfactory level was reached, further improvements had minimal impact. Therefore, the focus should be on achieving satisfactory grades. Other objective identity factors, such as age and gender, were less influential than academic performance. A web-based application incorporating the developed models was created, offering an open-access tool for forecasting dropout risks, with all code made publicly available for further research into undergraduate performance, which could be extended to other nations.

Keywords: undergraduate students; dropout risk assessment; graduation prediction model; educational data augmentation; machine learning classification

check for updates

Academic Editor: Sikha Bagui

Received: 17 August 2025
Revised: 25 September 2025
Accepted: 28 September 2025
Published: 23 October 2025

Citation: Wang, S.; He, J. Evaluating and Forecasting Undergraduate Dropouts Using Machine Learning for Domestic and International Students. *Technologies* **2025**, *13*, 480. https://doi.org/10.3390/technologies13110480

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Critical Need to Evaluate Dropouts

Reducing undergraduate dropout rates is a significant challenge for higher education institutions worldwide [1]. The consequences extend beyond individual students, affecting their personal development and career prospects and contributing to broader societal and economic instability [2]. However, dropout is not universally detrimental. In the context of labor market transformation, some students who leave academic education pursue vocational and technical education, which in many countries (including Europe) is gaining importance by providing stable employment and competitive earnings, often more quickly

than academic paths [3]. Thus, dropout may, in certain cases, represent an adaptation to labor market realities rather than a failure.

High dropout rates underline the stability and quality of educational systems, disrupt social cohesion, and pose challenges for economic growth [4]. At the same time, dropout reflects deeper social and cultural changes, such as the depreciation of academic diplomas due to declining educational quality, the rising demand for skills in the labor market, and the growing role of vocational education. In this context, dropout may indicate both a loss of certain educational benefits (e.g., social mobility, degree-related opportunities) and a rational choice for quicker labor market entry and economic independence [5].

Despite widespread recognition of the issue, there is no universally accepted definition of dropout. Variations in how dropout is defined, along with differences in data sources and calculation methods, result in inconsistent estimates of dropout rates [6]. These discrepancies complicate the development of a cohesive global strategy and underscore the need to understand the socio-economic and cultural drivers of dropout. Therefore, further research into the causes and predictors of undergraduate dropout is crucial for creating effective interventions and improving student retention in higher education.

While dropout risks affect all students, domestic and international students face distinct challenges that may influence their likelihood of leaving university [7,8]. Domestic students often have different academic profiles and experiences compared to international students, suggesting the need for studies that address their specific risks [8]. Meanwhile, international students face additional obstacles, such as language barriers, cultural adjustment, and financial pressures, which can impact their academic performance and overall university experience [9,10]. These challenges may contribute to higher dropout rates within this group and require separate investigation to develop targeted strategies for retention.

For example, a study by Smith et al. [11] at a Canadian university found that international students relied more on social networks, engaged more with campus life, and had a stronger connection to faith compared to domestic students. They also reported a greater need for a safe living environment and access to healthy food, factors that significantly impacted their academic success. Exploring these distinct factors will help universities create more effective retention strategies that address the unique needs of each group.

1.2. Current State of Literature

Undergraduate dropout risk assessment is a critical issue in higher education. Research consistently demonstrates that both subjective factors, such as mental health and motivation, and objective factors, including academic performance and financial status, significantly influence dropout intention [12–17]. For instance, Savio et al. [12] tracked students before and after the COVID-19 pandemic, revealing a dramatic surge in dropout intention, from 0% to 25–40%, directly linked to deteriorating mental health and financial instability, highlighting the profound effect of external crises.

Cultural and technological factors also play an important role in shaping dropout risk. Migration patterns, content consumption habits, and the pervasive influence of social media are all shifting students' priorities and behaviors, potentially influencing their academic engagement and decision to persist or dropout. For instance, Loureiro et al. [13] show that while affirmative action policies increased enrollment of Black, mixed-race, Indigenous, and quilombola students, dropout rates for these groups also rose, suggesting that cultural factors continue to heighten dropout risk and that multicultural policy integration remains only partially successful.

Crucially, the drivers of dropout risk differ markedly between domestic and international student populations. Anastasiou et al. [14], using an International Students' Expectation Risk Assessment Matrix, found that international students prioritize campus living conditions and social integration as critical factors influencing their persistence. In contrast, Costa et al. [15], employing Kaplan–Meier estimation and Cox modeling, identified strong academic performance as a primary protective factor against dropout overall, but emphasized that international students face compounded risks specifically stemming from language barriers, cultural adjustment, and integration challenges within the technological and social media-driven academic environment. These distinct risk profiles underscore the necessity for developing group-specific predictive models that account for both socio-cultural and technological dimensions.

Methodologically, traditional approaches to assessing dropout risk have often relied on resource-intensive longitudinal surveys, questionnaires, or interviews [16–20]. Llauro et al. [16], for example, longitudinally assessed instructors' ability to predict student dropout intentions. While demonstrating moderate accuracy, this approach was found to demand an unsustainable administrative workload, highlighting the critical need for scalable, automated solutions leveraging artificial intelligence technologies.

Machine learning (ML) and deep learning (DL) techniques offer promising alternatives by effectively modeling complex, nonlinear interactions between diverse predictive factors [21–24]. ML models have successfully established links, such as higher family income correlating with lower dropout risk [22], or identified gender as a key predictor in online education contexts using algorithms like Light Gradient Boosting Machine [23]. More advanced DL methods, particularly the Conditional Tabular Generative Adversarial Network (CTGAN) technique based on neural networks, further address the challenges of limited data and potential overfitting [25–27]. In combination with ML models, CTGAN enhances the robustness of predictions by leveraging the complementary strengths of data augmentation and predictive modeling [21].

Despite significant advances, limitations persist in the current literature. Most studies focus on isolated factors, such as mental health or academic pressure, without assessing their relative weights and complex interactions. The differential impact of subjective and objective factors on domestic versus international students is also poorly understood [14,24,28]. Additionally, while ML and CTGAN show potential, their application in modeling the complex, interdependent pathways leading to dropout, especially across student groups, remains underdeveloped. Therefore, research is needed that uses advanced computational methods, like deep learning, to assess the interplay of multiple risk factors, compare predictive models, identify key drivers for different student cohorts, and quantify the correlations between these factors to guide more targeted interventions.

1.3. Contributions of the Present Research

In this study, advanced ML and CTGAN techniques were applied to evaluate and predict undergraduate dropout risks, aiming to reveal the complex relationships between various factors and the outcome, as well as to develop a predictive tool. The workflow for conducting the ML research is depicted in Figure 1.

The dataset used for this research includes 3544 data points for domestic students and 86 data points for international students, with 23 input features and 1 target outcome: whether a student will graduate or dropout. For domestic students, 85% of the data was used for training, while 15% was reserved for testing. For international students, data augmentation was performed using the CTGAN technique, a deep learning-based approach, to generate 487 synthetic data points. The training was performed using cross-validation on the synthetic data, while testing was conducted on the real 86 data points to

ensure model reliability and generalizability. The Automated Machine Learning (AutoML) approach was used to streamline the process, offering a low-code solution accessible to researchers across various fields. Key factors influencing dropout risk were further assessed using SHapley Additive exPlanations (SHAP), and a web-based GUI app was developed using the Streamlit cloud platform for easy access.

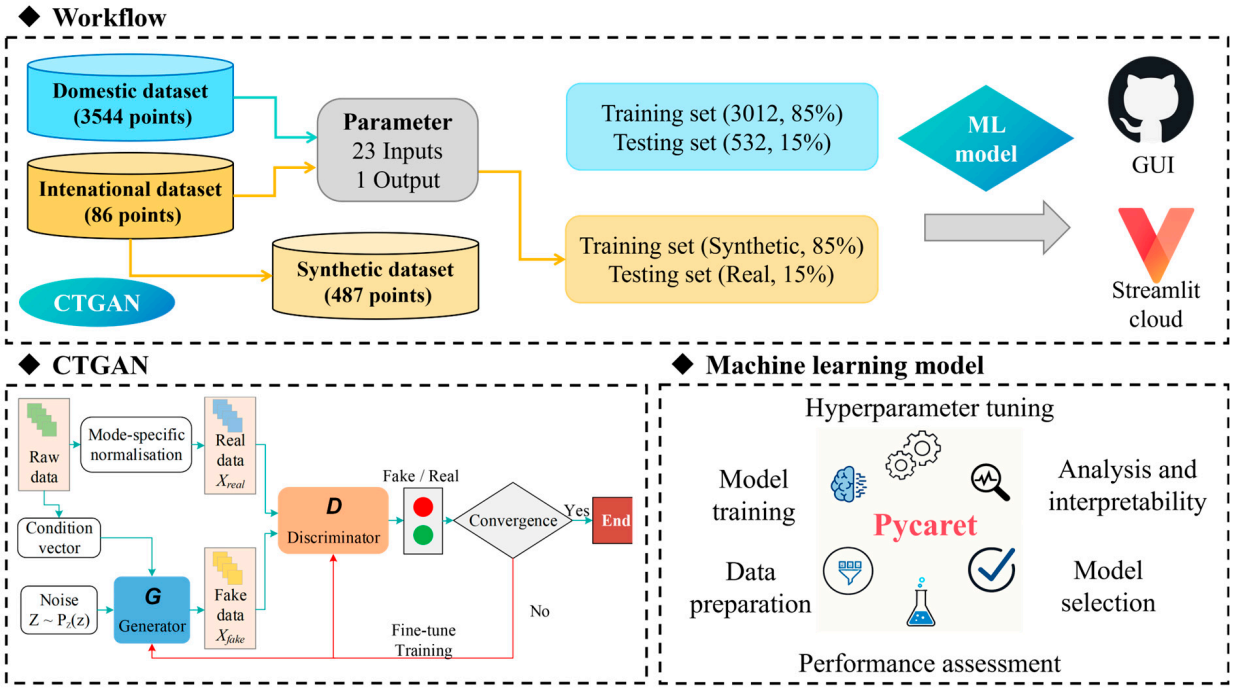


Figure 1. Workflow for performing machine learning research.

The study addresses three key research questions:

(1) What are the relationships between the various factors and the dropout risk for both domestic and international students?

(2) Do the most significant factors differ between domestic and international students, and are objective factors more influential than subjective factors?

(3) How can educators use the forecasting models in practice, and how can researchers extend this study to other nations?

To answer the first question, a comprehensive dataset and data augmentation techniques were used to explore complex relationships between factors and dropout risk through ML. For the second question, separate ML models were developed for domestic and international students, combined with Explainable Artificial Intelligence (XAI) techniques like SHAP to rank impact factors and identify key ones for tailored support. For the third question, an online, web-based app was created for mobile access, with all ML code open-access in the Supplementary Materials for further adaptation to other nations.

2. Machine Learning Methodology

A data augmentation strategy, in conjunction with an AutoML framework, was utilized to develop robust data-driven predictive models for undergraduate dropouts. To assess and interpret the resulting ML models, a variety of classification evaluation metrics and XAI techniques were employed. The foundational background supporting the autonomy of this study is briefly outlined below, with further details available in the referenced literature.

2.1. Data Augmentation for Enhancing Models

Building robust, data-driven ML models in real-world scenarios is often hindered by limited data availability. Additional data collection surveys are often resource-intensive and costly, rendering them an impractical solution. Consequently, deep learning data augmentation techniques like CTGAN are becoming increasingly important for generating high-quality synthetic data [29].

CTGAN, introduced by Xu et al. in 2019 [30], is a novel Generative Adversarial Network (GAN) variant specifically designed for generating synthetic tabular data, and falls within the domain of deep learning. It effectively preserves the statistical properties of the original dataset, making it highly suitable for data augmentation. The architecture consists of two neural networks trained in an adversarial manner: a Generator (G), which creates synthetic data samples, and a Discriminator (D), which attempts to distinguish between real data and the synthetic data produced by the generator (refer to Figure 1). Although G and D continuously compete without fully achieving their individual goals during this adversarial process, the G neural network gradually learns to produce realistic synthetic data that captures meaningful statistical characteristics from the original dataset [29]. This ability to generate high-quality synthetic tabular data is the core purpose of applying CTGAN.

2.2. Automated Machine Learning Models

AutoML is transforming the field by enhancing accessibility and efficiency. It accelerates development and reduces the need for extensive coding expertise, democratizing ML technologies [31–33].

PyCaret, an open-source library licenced under the MIT License, exemplifies this approach. It serves as a Python 3.7 wrapper that integrates multiple libraries and frameworks, such as scikit-learn and various ensemble models. This integration enables diverse tuning strategies and model customization (See Figure 1).

PyCaret further simplifies model evaluation by providing a concise, tabular summarized overview of key performance metrics and training times across models. This facilitates rapid comparison and informed selection for specific tasks. Including training durations also offers critical insights into computational demands, aiding scalability assessment and resource planning. Finally, PyCaret supports flexible deployment to cloud platforms or local servers, ensuring seamless integration into production environments and enhancing operational readiness [32].

2.3. Classification Evaluation Metrics

Performance metrics are crucial for rigorously assessing the effectiveness of classification models in PyCaret, such as distinguishing between graduate and dropout outcomes. These metrics quantify various aspects of model performance, such as accuracy, reliability, and its ability to differentiate between classes. This enables robust model comparisons and informs targeted improvements. This study uses several key metrics [33,34]:

- Accuracy: Proportion of correct predictions.
- AUC (Area Under the Curve): Measures class separation ability.
- Recall: Ability to correctly identify positive cases.
- Precision: Reliability of positive predictions.
- F1 Score: Harmonic mean of precision and recall, useful for imbalanced classes.
- Cohen's Kappa: Agreement between predictions and actual outcomes, beyond chance.

Each metric ranges from 0 to 1, with 1 indicating optimal performance [35]. Together, these metrics offer a comprehensive view of model strengths and weaknesses.

2.4. Interpretability and Model Explainability

As ML models grow more complex, their interpretability becomes essential for trust and accountability. XAI techniques address this by making model behavior transparent. A key method is SHAP, based on game theory. SHAP quantifies each feature's contribution to a specific prediction using Shapley values, capturing both individual impacts and feature interactions by treating them as collaborators in a cooperative game, offering detailed insight into individual predictions [35].

Complementing SHAP, ICE plots and PDPs help understand feature influence. ICE plots show how a prediction for a single instance change as one feature varies, revealing instance-specific effects. PDPs illustrate the average effect of a feature across the dataset, marginalizing over others, offering a global view [36]. Together, these techniques provide crucial insights into model decision-making, significantly enhancing interpretability and trustworthiness.

3. Dataset Description

The real-world data sources utilized in this study include data from multiple institutions, such as the annual data from the General Directorate of Higher Education and data from the Contemporary Portugal Database [37]. Due to the limited availability of data for international students, the CTGAN was meticulously trained to augment the dataset by generating synthetic data, thus enhancing the analysis.

3.1. Real-World Dataset

The dataset comprises 3544 records for domestic students and 86 records for international students, capturing 23 features that encompass various aspects of student identity, academic performance, and socioeconomic status. These features provide a comprehensive insight into student data.

An analysis of the summary statistics in Table 1 reveals notable variation across the features, with some showing non-normal distributions. For instance, the Marital and Need features display significant skewness and kurtosis, indicating highly skewed distributions. Similarly, the target variable, representing the outcome of interest, exhibits a skewness of 4.34 and kurtosis of 20.89, further suggesting a non-normal distribution. While most features generally adhere to expected distribution patterns, the observed skewness in several variables may introduce challenges for modeling. Nevertheless, these findings provide valuable insights, and the work remains crucial for advancing our understanding of the dataset, with potential avenues for refinement and further exploration.

3.2. Synthetic Data Generation

Due to the limited size of the international student dataset (only 86 real samples), we applied the CTGAN-based data augmentation method (Section 2.1) to generate a complementary synthetic dataset. The well-trained CTGAN effectively captured the underlying distribution of the original data, generating a synthetic dataset with statistically meaningful properties. Both the real and synthetic datasets are available in the Supplementary Material; key statistics of the synthetic dataset are summarized in Table 2.

Figure 2 compares the density distributions of the real and synthetic datasets, showing that the synthetic data closely follows the distribution of the real data across all variables, thereby confirming that CTGAN has effectively learned the underlying patterns. Notably, the proportions of discontinuous target outputs are almost identical in both datasets.

Figure 3 presents correlation matrices for both datasets. The Pearson correlation coefficient (ρ) ranges from -1 to 1, indicating strong positive (near 1), strong negative (near

−1), or weak/no linear correlation (near 0). Key relationships, such as those between the Target and the 1st/2nd semester approved courses, show strong correlations ($\rho > 0.55$).

Table 1. Key distributional indicators for the real-world dataset.

Data	Abbreviation	Description	Mean	Std	Min	Max	Skewness	Kurtosis
Inputs	Marital	Marital status	1.18	0.61	1	6	4.34	20.89
	Mode	Application mode	6.81	5.25	1	18	0.13	−1.50
	Order	Application order	1.75	1.33	1	6	1.82	2.30
	Course	Course type	9.93	4.34	1	17	−0.23	−0.98
	Attendance	Daytime/evening attendance	0.89	0.32	0	1	−2.45	4.02
	Qualification	Previous qualification	2.55	3.95	1	17	2.38	3.98
	Nationality	Nationality	1.24	1.70	1	21	7.43	56.24
	Mother-Q	Mother qualification	12.56	9.01	1	29	−0.15	−1.63
	Father-Q	Father qualification	16.66	10.99	1	34	−0.33	−1.54
	Mother-O	Mother occupation	7.21	3.71	1	32	1.84	9.75
	Father-O	Father occupation	7.65	4.32	1	46	3.89	28.67
	Displaced	Displaced	0.55	0.50	0	1	−0.20	−1.96
	Need	Educational special need	0.01	0.10	0	1	9.37	85.76
	Debtor	Debtor	0.11	0.32	0	1	2.43	3.92
	Fee	Tuition fee	0.87	0.34	0	1	−2.15	2.62
	Gender	Gender	0.34	0.48	0	1	0.66	−1.57
	Scholarship	Scholarship	0.27	0.44	0	1	1.05	−0.89
	Age	Age	23.46	7.83	17	70	1.99	3.80
	1st	1st semester approved course	4.79	3.24	0	26	0.75	2.86
	2nd	2nd semester approved course	4.52	3.16	0	20	0.27	0.66
	Unemployment	Unemployment rate	11.63	2.67	7.6	16.2	0.19	−1.00
	Inflation	Inflation rate	1.23	1.38	−0.8	3.7	0.27	−1.04
	GDP	GDP	−0.01	2.26	−4.06	3.51	−0.39	−1.00
Output	Target	Target	0.61	0.49	0	1	4.34	20.89

Table 2. Key distributional indicators for the CTGAN-synthetic dataset.

Data	Abbreviation	Mean	Std	Min	Max	Skewness	Kurtosis
Inputs	Marital	1.07	0.36	1	4	4.88	22.70
	Mode	6.73	4.87	1	18	0.51	−1.01
	Order	1.47	0.77	1	5	1.32	0.25
	Course	11.44	4.26	1	17	−0.42	−0.91
	Attendance	0.95	0.21	0	1	−4.27	16.22
	Qualification	1.67	2.72	1	14	3.90	13.45
	Nationality	12.56	3.82	2	21	−0.73	0.28
	Mother-Q	11.11	8.45	1	28	0.14	−1.68
	Father-Q	10.53	10.48	1	28	0.70	−1.20
	Mother-O	7.46	3.06	1	25	−0.43	−1.00
	Father-O	7.56	2.40	1	26	−0.09	−0.18
	Displaced	0.53	0.50	0	1	−0.11	−1.99
	Need	0.004	0.064	0	1	15.51	238.50
	Debtor	0.27	0.44	0	1	1.03	−0.94
	Fee	0.75	0.43	0	1	−1.15	−0.67
	Gender	0.27	0.44	0	1	1.04	−0.91
	Scholarship	0.21	0.41	0	1	1.44	0.08
	Age	22.90	6.56	18	59	2.64	9.00
	1st	5.46	4.04	0	18	1.21	2.25
	2nd	4.68	3.17	0	12	−0.02	−0.62
	Unemployment	12.13	2.32	7.6	15.33	0.07	−0.88
	Inflation	1.10	1.19	−0.8	3.7	0.45	−0.68
	GDP	0.018	2.41	−4.06	3.51	−0.20	−1.22
Output	Target	0.66	0.48	0	1	−0.66	−1.56

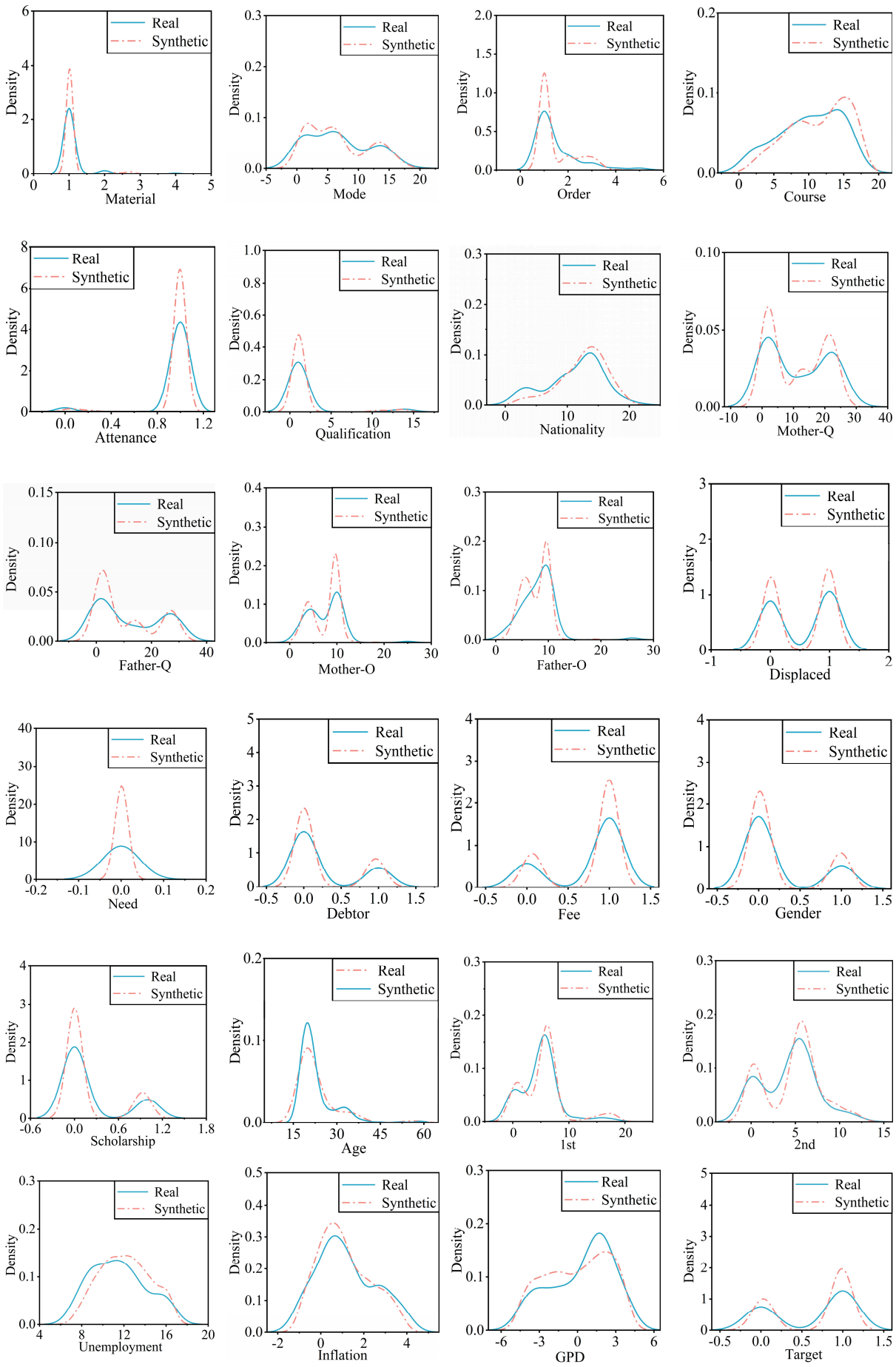
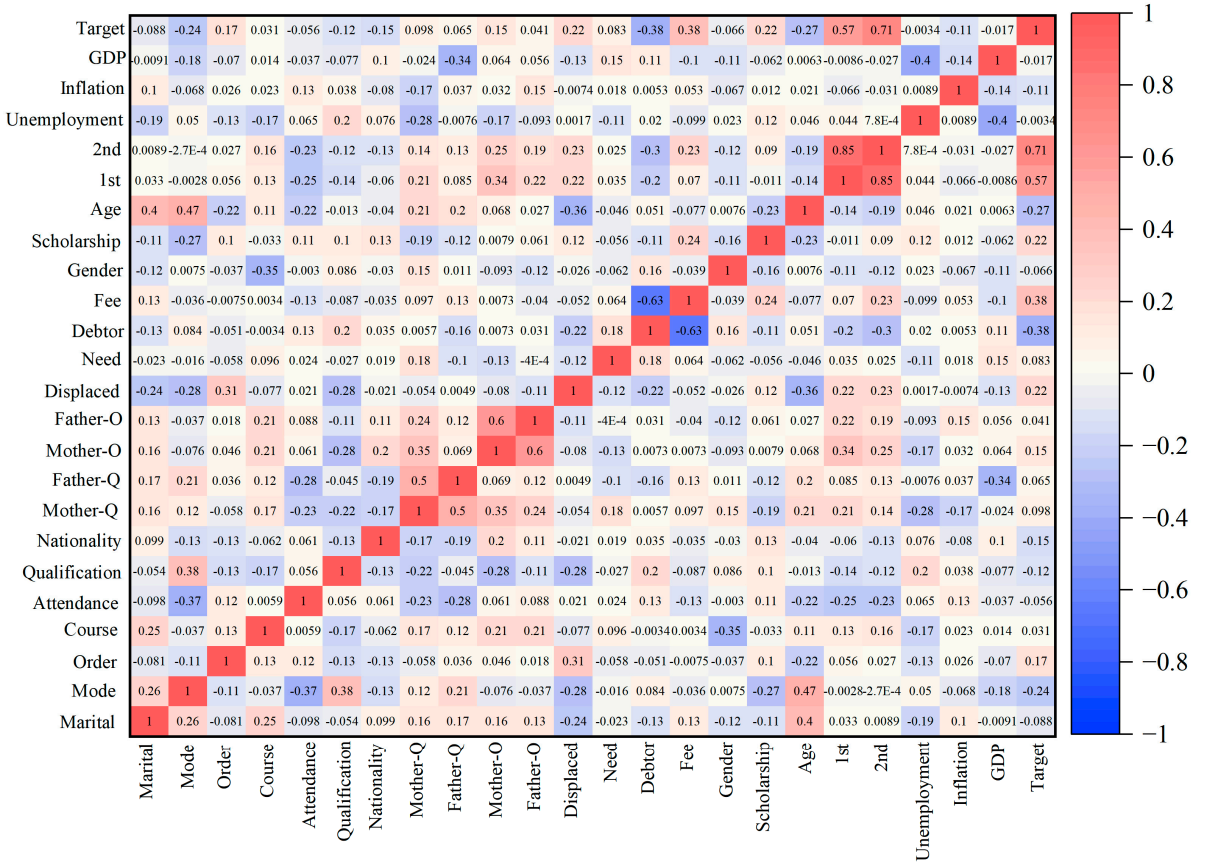
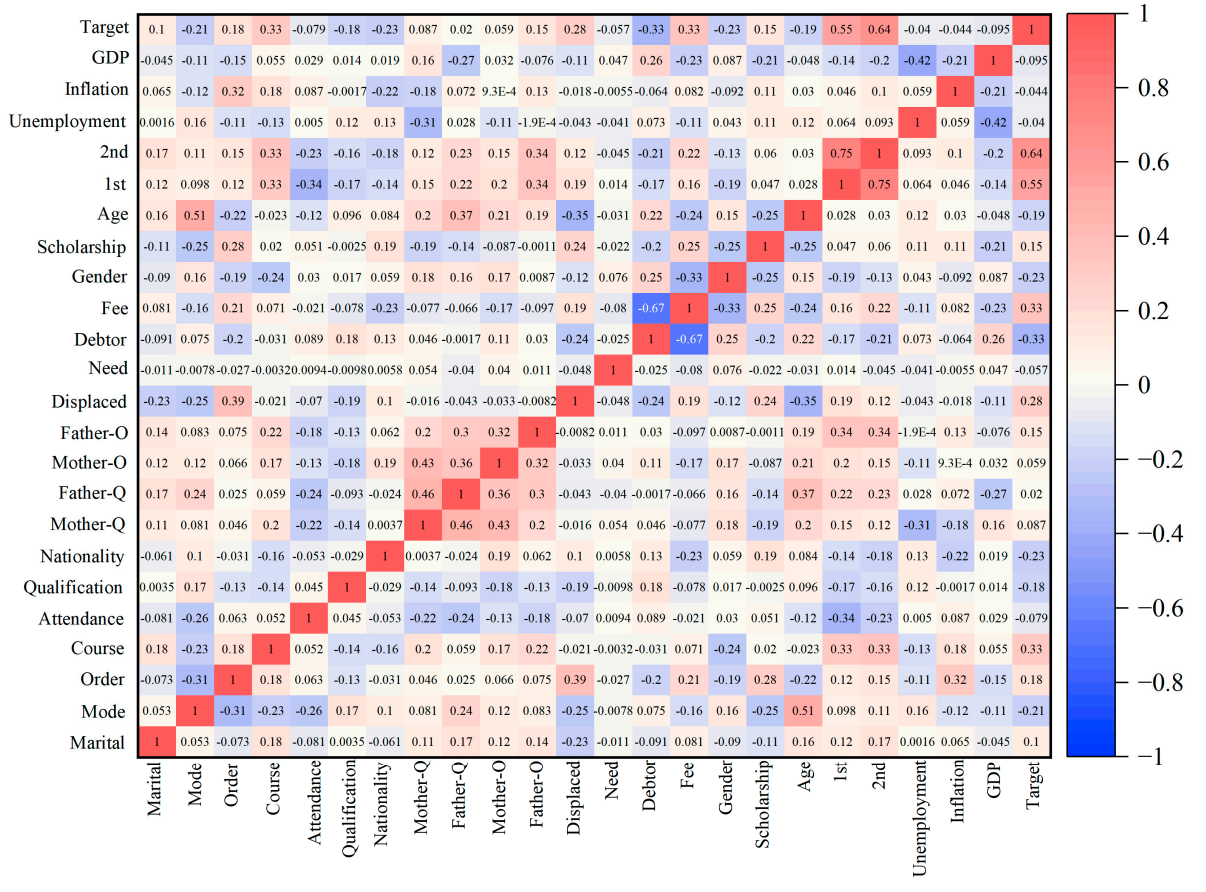


Figure 2. Visualization of real versus synthetic data distributions.



(**a**)



(**b**)

Figure 3. Correlation analysis of real versus synthetic datasets. (**a**) Real dataset. (**b**) Synthetic dataset.

Table 3 illustrates the distributional similarity between the two datasets. The Jensen–Shannon Divergence (JSD) ranges from 0 to 1, where values closer to 0 indicate stronger distributional similarity, and values closer to 1 indicate greater divergence. Several key features, such as scholarship and gender, exhibit strong distributional similarity [38].

Table 3. Key distributional indicators for real dataset and CTGAN-synthetic dataset.

Data	Abbreviation	Real Data Mean	Real Data Std	Synthetic Data Mean	Synthetic Data Std	JSD
Inputs	Marital	1.08	0.38	1.07	0.36	0.034
	Mode	6.74	4.98	6.73	4.87	0.238
	Order	1.49	0.92	1.47	0.77	0.046
	Course	10.11	4.41	11.44	4.26	0.067
	Attendance	0.95	0.21	0.95	0.21	0.008
	Qualification	1.74	2.97	1.67	2.72	0.044
	Nationality	11.23	4.47	12.56	3.82	0.693
	Mother-Q	11.40	9.61	11.11	8.45	0.206
	Father-Q	11.78	11.46	10.53	10.48	0.152
	Mother-O	7.5	3.65	7.46	3.06	0.140
	Father-O	8.01	3.19	7.56	2.40	0.038
	Displaced	0.55	0.50	0.53	0.50	0.0002
	Need	0.12	0.11	0.004	0.064	0.001
	Debtor	0.26	0.44	0.27	0.44	0.021
	Fee	0.74	0.44	0.75	0.43	0.011
	Gender	0.24	0.43	0.27	0.44	0.004
	Scholarship	0.21	0.41	0.21	0.41	0.003
	Age	23.09	7.31	22.90	6.56	0.035
	1st	4.91	3.43	5.46	4.04	0.058
	2nd	4.31	3.04	4.68	3.17	0.042
	Unemployment	11.49	2.55	12.13	2.32	0.295
	Inflation	1.18	1.33	1.10	1.19	0.328
	GDP	0.34	2.37	0.018	2.41	0.284
Output	Target	0.63	0.49	0.66	0.48	0.001

While these plots provide valuable insights into potential variable influences, it is important to note that correlation values can vary significantly depending on the well-balanced distribution of the dataset [29]. Therefore, Section 4.3 provides a more comprehensive analysis of feature importance using the SHAP method.

4. Results

Leveraging machine learning approaches, this research investigates undergraduate dropout for domestic and international students. The implementation code is provided in the Supplementary Material to enable replication and further application. This section describes the modeling process, performance evaluation, and interpretation of the prediction results.

4.1. Development of the Graduation Prediction Model

Classification algorithms were applied using an 85% training and 15% testing split, as illustrated in Figure 4. The synthetic dataset served exclusively as the training set for the international student model, resulting in a validation strategy of cross-validation on synthetic data and a single test on real data. The 10-fold cross-validation approach maximized training data utilization and mitigated overfitting risks. Additionally, a real data comparison model is introduced in the CTGAN model for international students, enabling validation of the results with actual data.

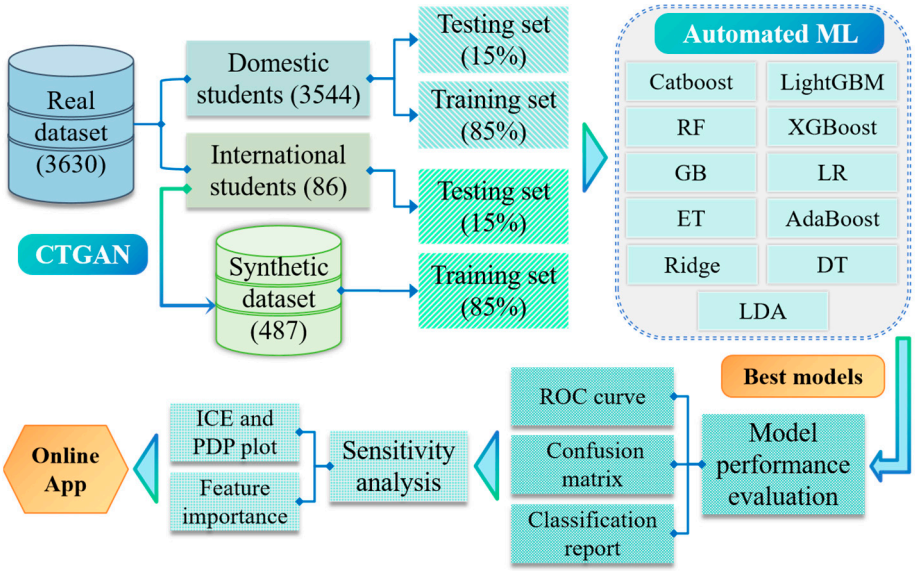


Figure 4. Synthetic data-augmented and machine learning assessment approaches.

Tables 4 and 5 present a detailed performance comparison of the top 10 models used for predicting dropout rates among domestic and international students, respectively. These models, developed with widely adopted algorithms, were evaluated based on the performance metrics outlined in Section 2.3.

Table 4. Model performance comparison for forecasting dropout of domestic students.

Abbr.	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
LightGBM	Light Gradient Boosting Machine	0.9014	0.9445	0.9014	0.9037	0.9001	0.7887
XGBoost	Extreme Gradient Boosting	0.9004	0.9433	0.9004	0.9019	0.8992	0.7869
RF	Random Forest Classifier	0.9001	0.9471	0.9001	0.9016	0.8989	0.7864
GB	Gradient Boosting Classifier	0.9001	0.9469	0.9001	0.9022	0.8986	0.7855
ET	Extra Trees Classifier	0.8997	0.9479	0.8997	0.9011	0.8986	0.7855
AdaBoost	Ada Boost Classifier	0.8981	0.9394	0.8981	0.8990	0.8970	0.7825
LR	Logistic Regression	0.8858	0.9307	0.8858	0.8859	0.8850	0.7576
Ridge	Ridge Classifier	0.8848	0.9302	0.8848	0.8857	0.8836	0.7540
LDA	Linear Discriminant Analysis	0.8845	0.9301	0.8845	0.8853	0.8833	0.7534
DT	Decision Tree Classifier	0.8476	0.8415	0.8476	0.8486	0.8477	0.6808

Among the models tested, the ensemble Boosting classifier stood out with the strongest overall performance. Specifically, for domestic students, the Light Gradient Boosting Machine (LightGBM) achieved the highest accuracy at 0.9014, while for international students, the CatBoost Classifier (CatBoost) emerged as the top performer, attaining an accuracy of 0.9004. This comparison highlights the nuanced performance of different algorithms across diverse student populations, demonstrating the adaptability and effectiveness of machine learning techniques in predicting dropout risk.

To further assess the risks associated with applying the CTGAN method for international students, we adopted the Train on Real, Test on Real (TRTR) approach for validation. The CatBoost classifier achieved an accuracy of 0.8518, which was lower than the performance obtained under the Train on Synthetic, Test on Real (TSTR) setting. After applying the CTGAN method, however, the accuracy increased substantially to 0.9004, demonstrating that CTGAN effectively augmented the data and improved CatBoost's predictive capability on real data.

Table 5. Model performance comparison for forecasting dropout of international students.

Abbr.	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
Catboost	CatBoost Classifier	0.9004	0.9475	0.9004	0.9031	0.8988	0.7858
RF	Random Forest Classifier	0.8974	0.9395	0.9331	0.9108	0.9214	0.7735
GB	Gradient Boosting Classifier	0.8933	0.9400	0.9298	0.9077	0.9184	0.7640
ET	Extra Trees Classifier	0.8850	0.9373	0.9203	0.9037	0.9116	0.7470
Ridge	Ridge Classifier	0.8830	0.9301	0.9205	0.9017	0.9105	0.7414
LDA	Linear Discriminant Analysis	0.8830	0.9303	0.9205	0.9017	0.9105	0.7414
LightGBM	Light Gradient Boosting Machine	0.8810	0.9314	0.9138	0.9040	0.9081	0.7388
XGBoost	Extreme Gradient Boosting	0.8809	0.9340	0.9136	0.9042	0.9075	0.7393
AdaBoost	Ada Boost Classifier	0.8728	0.9324	0.9108	0.8959	0.9025	0.7188
LR	Logistic Regression	0.8727	0.9296	0.9140	0.8924	0.9026	0.7186

4.2. Performance Assessment of the Constructed Model

The performance of the LightGBM and CatBoost models was thoroughly evaluated to assess their prediction accuracy and provide insights for future research applications. The evaluation employed several visualization tools, including AUC plots, confusion matrices, and classification report heatmaps. Figure 5 illustrates the AUC curves for both domestic and international students, where the classes are numbered 0 and 1.

Class 0 represents the negative outcome (dropout), and Class 1 indicates the positive outcome (graduation). Both the LightGBM model for domestic students and the TSTR model for international students demonstrate strong classification performance, with similarly high true positive rates. Among them, the domestic model performs slightly better, which can be attributed to the larger dataset available for domestic students. The classification performance of the international students under the TRTR approach was generally moderate. After applying the TSTR method, the model classification performance improved.

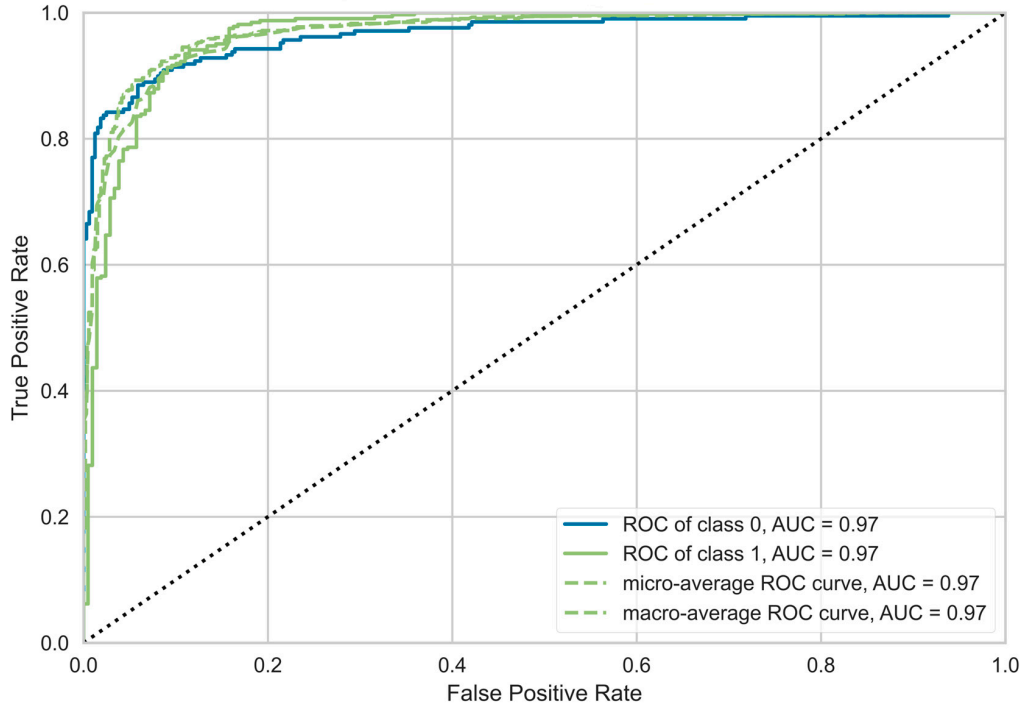
The confusion matrix, shown in Figure 6, visually represents misclassified instances. As indicated in Figure 5, the models perform with high accuracy, although some errors could be inevitable in machine learning applications, provided the models are not overfitted. Notably, misclassifications of dropouts as graduates could slightly outnumber the misclassification of graduates as dropouts for both domestic and international students. This could be an important consideration when deploying the models.

Figure 7 presents the classification report heatmap, which outlines the precision, recall, and F1 scores for each class. For domestic students, Class 1 (graduates) demonstrates high precision (0.912) and recall (0.954), yielding an F1 score of 0.922. For international students, Class 1 shows slightly lower precision (0.897), yet the model still demonstrates robust performance across all metrics, with high precision and recall values.

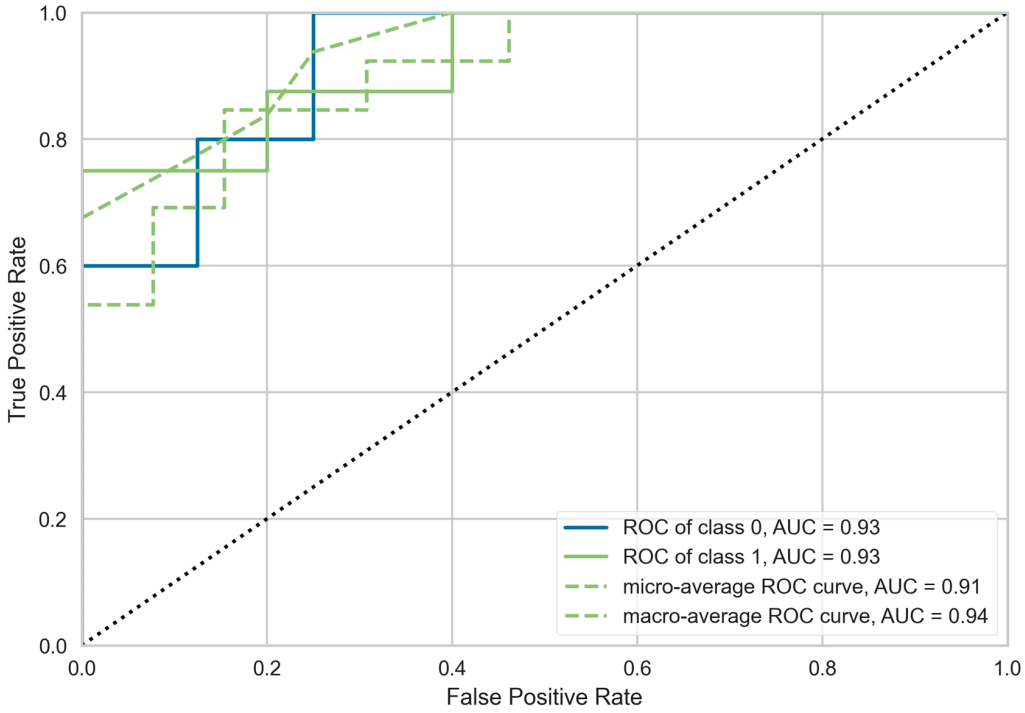
4.3. Analysis of Key Factors Influencing Dropout

SHAP values provide a detailed interpretation of feature importance, quantifying both the direction and magnitude of each variable's influence on the model's predictions. As illustrated in Figure 8, a color gradient ranging from blue to red represents the spectrum of SHAP values from lowest to highest, with each point corresponding to an individual observation within the dataset. The analysis reveals that subjective academic performance (the number of courses passed) is the primary determinant of undergraduate dropout risk. Notably, courses passed in the second semester (2nd) exert the strongest influence; higher pass rates correlate with significantly increased SHAP values, indicating a substantially

greater likelihood of graduation. A similar, though slightly less pronounced, trend is observed for courses passed in the first semester (1st).

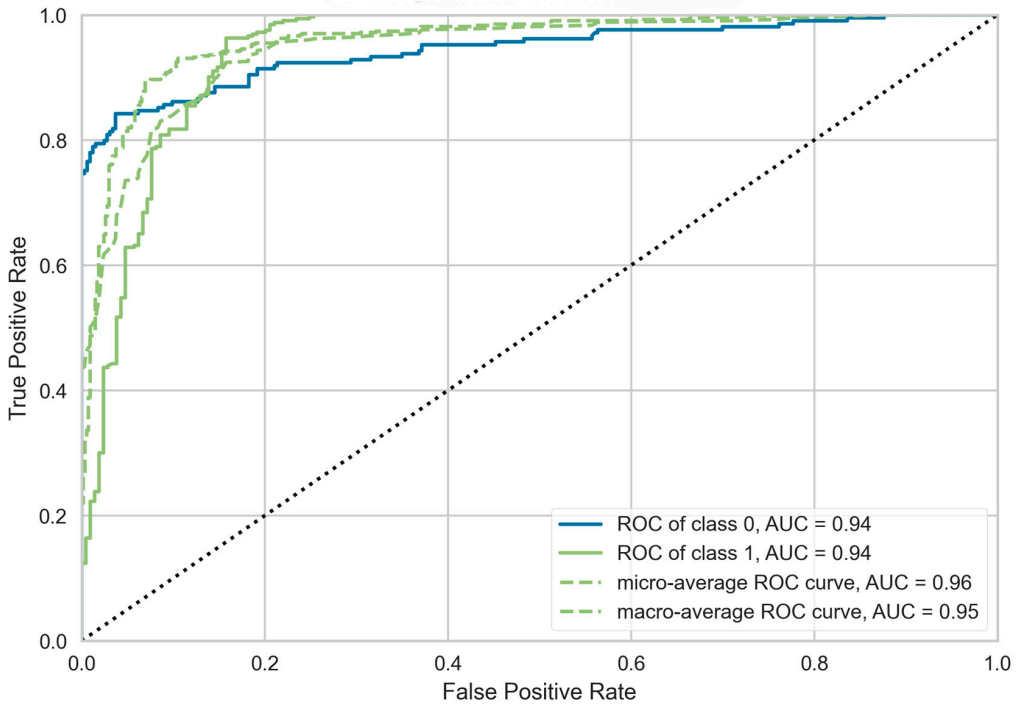


(**a**)



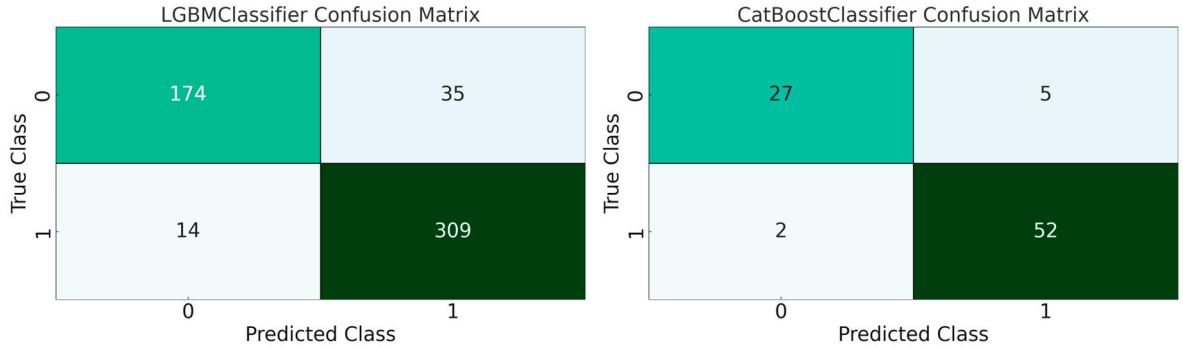
(**b**)

Figure 5. *Cont*.



(**c**)

Figure 5. Receiver operating characteristic curve for the constructed models. (**a**) Domestic student forecasting model. (**b**) International student forecasting model based on real data. (**c**) International student forecasting model based on synthetic data.

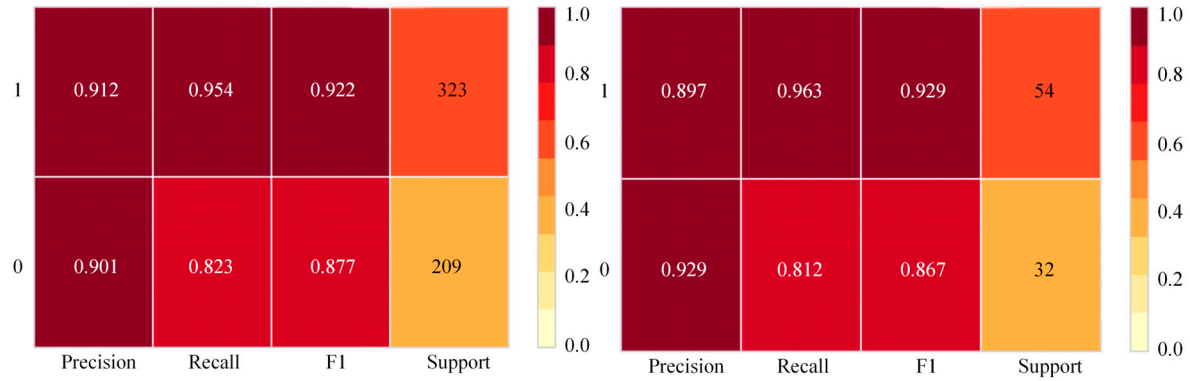


(**a**) (**b**)

Figure 6. Confusion matrix evaluation of the constructed models. (**a**) Domestic student forecasting model. (**b**) International student forecasting model.

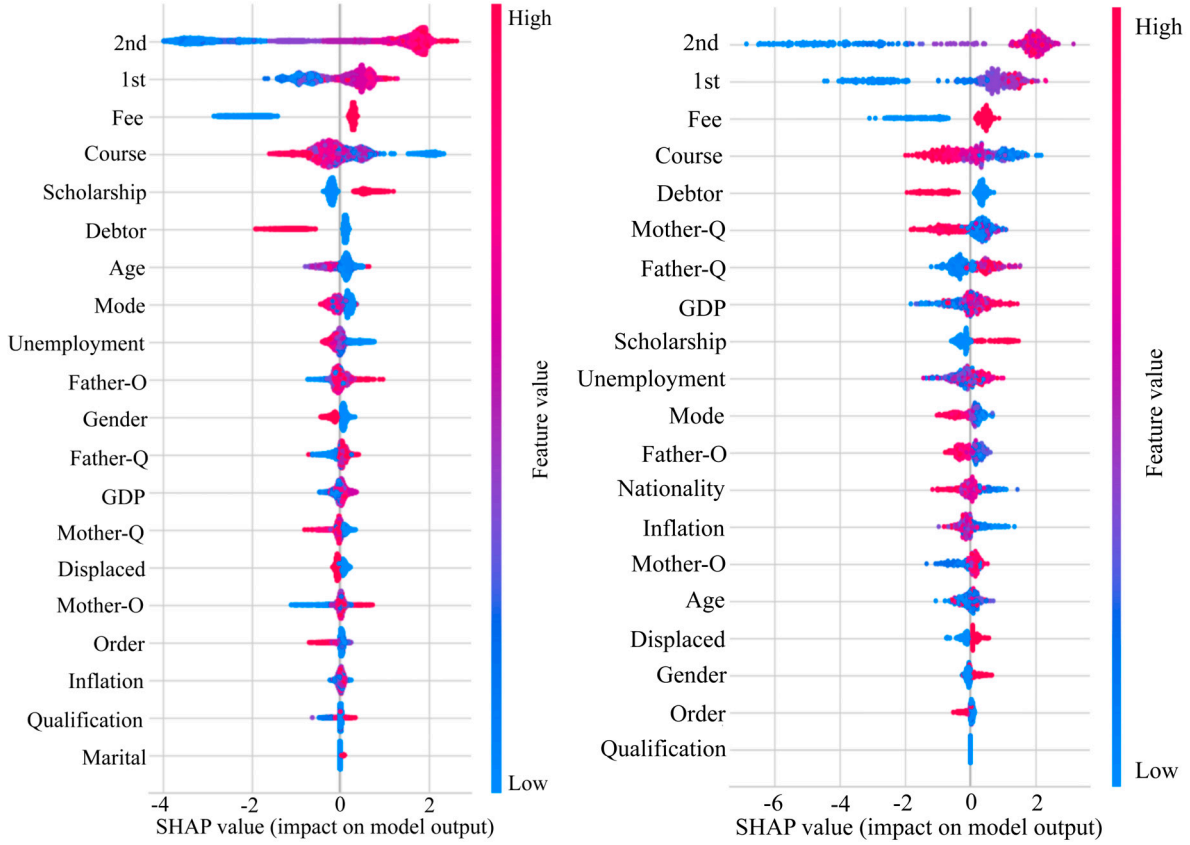
In contrast, objective demographic factors such as age, gender, family situation, and socioeconomic background demonstrate minimal predictive power, emphasizing that dropout propensity is predominantly linked to a student's academic engagement and performance rather than inherent characteristics. Additionally, the SHAP plot reveals that there is no significant difference between domestic and international students regarding the factors influencing dropout risk, as their top four most influential features are identical.

Complementary analysis using ICE and PDP in Figure 9 offers deeper insights into the relationship between the top features (2nd and 1st) and the predicted probability of dropout or graduation for both student cohorts. The ICE plots illuminate individual variations in response, while the PDP confirms the overall global trend identified by the SHAP analysis (see Figure 7).



(**a**) (**b**)

Figure 7. Combined assessment of model performance by multiple indicators. (**a**) Domestic student forecasting model. (**b**) International student forecasting model.



(**a**) (**b**)

Figure 8. Feature importance rankings in explainable AI. (**a**) Domestic student forecasting model. (**b**) International student forecasting model.

Together, these techniques validate the developed models, demonstrating their effectiveness in capturing the various factors that influence undergraduate graduation outcomes. This integrated analytical approach provides valuable insights for educators and institutions to design targeted support strategies.

The findings provide clear and actionable insights. To reduce dropout risk, it is essential to improve academic performance, particularly pass rates during the critical first and second semesters. This relationship demonstrates a clear non-linear pattern: once a satisfactory level of course completion is achieved, it offers substantial protection against dropout, but further improvements beyond this threshold lead to diminishing returns. As

a result, university support strategies should focus on assisting students in reaching this satisfactory performance level early in their academic journey, directing intervention efforts toward this pivotal stage rather than solely aiming for universal academic excellence.

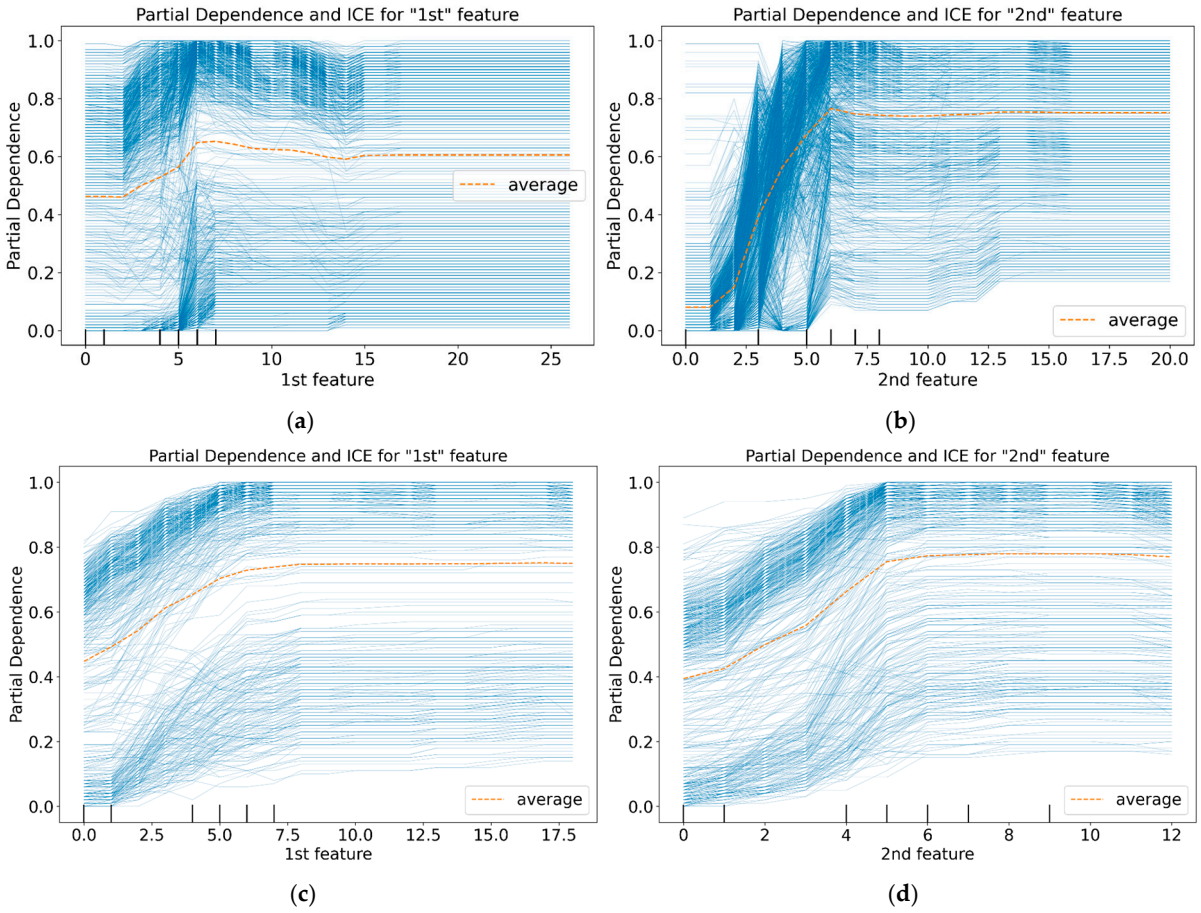


Figure 9. Trends in key factor impacts on outcomes using ICE and PDP plots. (**a**) Domestic student 1st. (**b**) Domestic student 2nd. (**c**) International student 1st. (**d**) International student 2nd.

4.4. Interactive Software Tool for Predictions

This web-based interactive application uses a GUI to forecast undergraduate dropout rates (accessible via the link: https://aaatu4bq3njwchzgunv4ph.streamlit.app/, accessed on 1 August 2025) for domestic and international students based on user-defined parameters. Developed on the open-source Streamlit platform the tool converts locally stored Python scripts and data files into a dynamic web application. Streamlit simplifies development, requiring minimal coding expertise and broadening accessibility for non-technical users.

The application's code and datasets are hosted on GitHub 3.15 and deployed via Streamlit Cloud, ensuring seamless online operation. This allows educators and researchers to utilize pre-trained models without additional programming. We have prepared the data for both domestic and international students along with the best-performing machine learning models, and have uploaded the code and files to GitHub. The repository is then detected and deployed, creating a unique web link for accessing the live Streamlit app. This link is distributed to researchers or the general public for easy access to the tool.

As illustrated in Figure 10, the interface features adjustable slider controls (left panel) for parameter input. Users modify these sliders within ranges reflecting the dataset scope (Section 3), and the integrated models generate real-time dropout predictions.

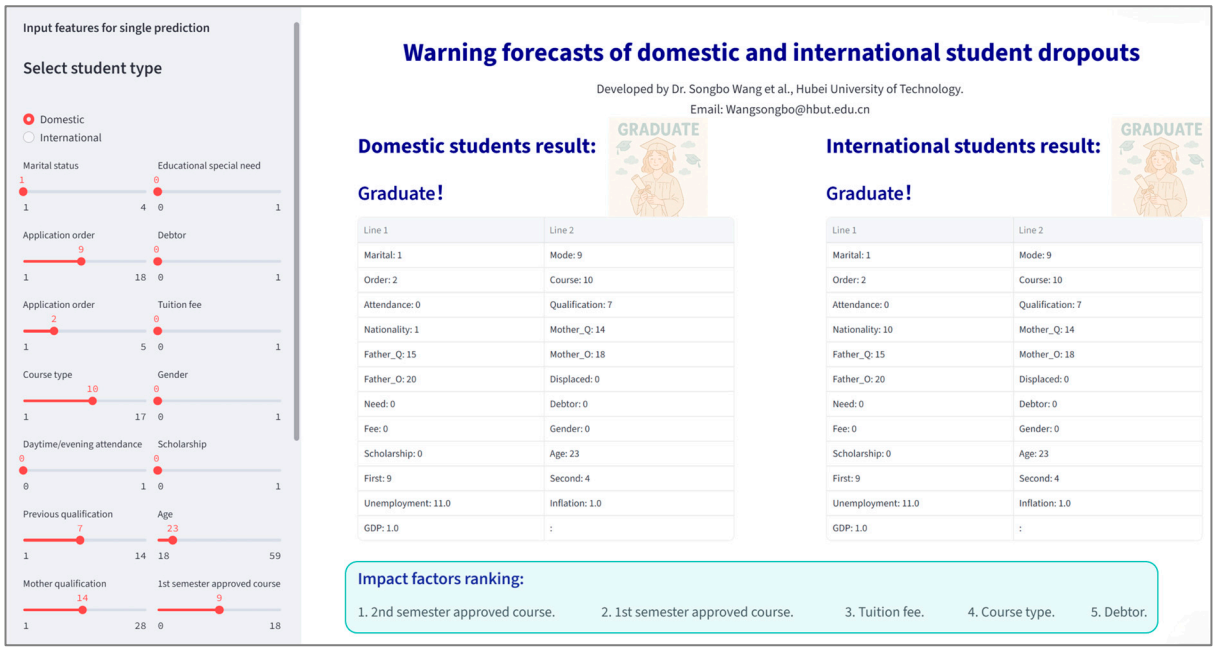


Figure 10. Implementation of early warning forecasts for dropouts using trained models.

5. Discussion

This study employed CTGAN to augment the international student dataset, addressing the challenge of limited sample size. While this approach enhanced the robustness and predictive power of the models, it also carries inherent risks. Because synthetic data are generated from small or skewed real samples, there is potential for amplifying noise, producing spurious patterns, or inducing overfitting. Moreover, discrepancies between synthetic and real data may undermine generalizability, and sociocultural complexities unique to international students are not easily replicated by generative models. These limitations suggest that although CTGAN provides a feasible means of overcoming data scarcity, its outcomes should be interpreted with caution and validated against larger and more diverse real-world datasets.

Importantly, the accuracy of the TSTR paradigm was further substantiated through integration with the TRTR model. To rigorously evaluate the representativeness of the synthetic dataset, we employed correlation matrices, Jensen–Shannon Divergence (JSD) statistics, and complementary validation techniques to compare it against the real dataset. By establishing a TRTR-based machine learning framework, we were able to conduct a systematic assessment of TSTR performance. This comprehensive validation enhances the methodological robustness of CTGAN and strengthens the credibility of its application in addressing data scarcity.

Both SHAP analysis and ICE/PDP plots confirm that academic performance in the first two semesters (specifically, the number of courses passed) is the strongest predictor of dropout risk. This aligns with prior literature emphasizing the critical transition period in the early stages of university life [15]. Notably, the relationship is non-linear: achieving a threshold of satisfactory grades (e.g., passing 4–6 courses per semester) drastically reduces the likelihood of dropout, but exceeding this threshold yields diminishing returns. This suggests that universities should prioritize helping students reach a baseline academic standard, rather than uniformly pushing for high excellence.

Contrary to some existing studies [12–17,23], objective factors (e.g., age, gender, parental education, nationality) showed minimal influence in our models. This may reflect Portugal's relatively homogeneous higher education demographics or the dominance of

academic performance in mediating other risks. For international students, expected stressors such as language barriers or financial need [14,22] were overshadowed by academic outcomes, implying that academic support could mitigate broader challenges.

Although academic performance remains the primary predictor of dropout risk, it is equally important to consider the broader socioeconomic and cultural factors that influence student retention. Factors such as financial stress, family background, and cultural adaptation, particularly for international students, often interact with academic outcomes [39–41]. These non-academic factors cannot be fully captured by academic performance metrics alone, yet they are crucial in understanding dropout risk.

An interdisciplinary approach, incorporating perspectives from sociology, economics, and psychology, can provide deeper insights into how these factors contribute to student attrition. By combining academic support with interventions addressing non-academic factors, universities can develop more comprehensive and effective strategies to assist at-risk students.

In practical terms, this study introduces a web-based application that provides real-time dropout predictions for students based on user-defined parameters. This tool holds promise for helping universities implement proactive measures to support at-risk students before they decide to leave their studies. However, it should be noted that despite the promising results, there are limitations to the current model, particularly in terms of its generalizability to other institutions and countries. Further research is required to validate these models across diverse educational settings and cultural contexts.

6. Conclusions

This study presents a novel approach to predicting undergraduate dropout risks using machine learning techniques, specifically tailored for both domestic and international students. The research highlights the critical role of academic performance, particularly in the first two semesters, in determining dropout likelihood, while also revealing that demographic and socioeconomic factors have a lesser impact on retention than previously believed.

By leveraging machine learning algorithms and data augmentation methods such as CTGAN, this research contributes significantly to the body of knowledge on student retention. The developed models, coupled with the web-based application, offer universities a valuable tool for forecasting dropout risks and implementing targeted interventions. The findings suggest that universities should focus on supporting students to reach a satisfactory academic level early in their studies, as this has the most profound effect on preventing dropouts.

However, when using the CTGAN method for data augmentation, it is crucial to pay special attention to risk assessment, such as the use of Jensen–Shannon Divergence (JSD), and compare the generated CTGAN data with the real data to ensure that the augmented data does not deviate from the actual distribution. Although CTGAN can generate high-quality synthetic data, if the generated data significantly differs from the distribution of the real data, it may negatively impact the model's predictive performance. Therefore, when applying CTGAN to generate augmented data, it is essential to rigorously evaluate the quality of the generated data to ensure that it is representative and accurately reflects the diversity of the student population. Such careful handling will help enhance the reliability of the model and ensure the practical applicability of the prediction results.

Looking ahead, the study's approach can be extended to other higher education systems across different countries to improve dropout prediction accuracy. Furthermore, the integration of the developed predictive models into university systems could pave the way for more personalized, data-driven strategies that can enhance student retention and

success. Future work should also explore the integration of additional features, such as psychological and social factors, which may provide a more comprehensive understanding of the dropout phenomenon.

The contribution of this research lies not only in its predictive capabilities but also in its practical application through an open-access, easy-to-use tool that offers potential benefits to both educators and policymakers in their efforts to reduce dropout rates and improve student outcomes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/technologies13110480/s1.

Author Contributions: S.W.: Conceptualization, Methodology, Formal analysis, Investigation, Writing—original draft, Funding acquisition. J.H.: Visualization, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Teaching Reform Research Project of Hubei Provincial Department of Education grant number 2024301.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments: This paper is funded by the Teaching Reform Research Project of Hubei Provincial Department of Education (Project No. 2024301).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roberts, M.; Shah, N.; Mali, D.; Hong, J. Dropout rates in undergraduate accounting degree programmes at a UK university: A 10-year longitudinal study. *Int. J. Manag. Educ.* **2025**, *23*, 101253. [CrossRef]
2. Bäulke, L.; Grunschel, C.; Dresel, M. Student dropout at university: A phase-orientated view on quitting studies and changing majors. *Eur. J. Psychol. Educ.* **2022**, *37*, 853–876. [CrossRef]
3. Belloc, F.; Maruotti, A.; Petrella, L. University drop-out: An Italian experience. *High Educ.* **2010**, *60*, 127–138. [CrossRef]
4. Lorenzo-Quiles, O.; Galdón-López, S.; Lendínez-Turón, A. Dropout at university. Variables involved on it. *Front. Educ.* **2023**, *8*, 1159864. [CrossRef]
5. Venuleo, C.; Mossi, P.; Salvatore, S. Educational subculture and dropping out in higher education: A longitudinal case study. *Stud. High. Educ.* **2016**, *41*, 321–342. [CrossRef]
6. Zhou, Y.; Zhao, J.; Zhang, J. Prediction of learners' dropout in E-learning based on the unusual behaviors. *Interact. Learn. Environ.* **2023**, *31*, 1796–1820. [CrossRef]
7. Glass, C.R.; Westmont, C.M. Comparative effects of belongingness on the academic success and cross-cultural interactions of domestic and international students. *Int. J. Intercult. Relat.* **2014**, *38*, 106–119. [CrossRef]
8. Jiang, X.; Xu, J.; Xu, X. An overview of domestic and international applications of digital technology in teaching in vocational education: Systematic literature mapping. *Educ. Inf. Technol.* **2024**, *29*, 16867–16899. [CrossRef]
9. Chiang, F.K.; Brooks, D.C.; Chen, H. Cross-cultural social contexts: A comparison of Chinese and US students' experiences in active learning classrooms. *Interact. Learn. Environ.* **2023**, *31*, 1623–1635. [CrossRef]
10. Sin, S.C.J.; Kim, K.S. How Are We the Same or Different: Information Needs and Barriers of Domestic and International Students. *J. Acad. Libr.* **2018**, *44*, 712–723. [CrossRef]
11. Smith, S.M.; Carter-Rogers, K.; Norris, M.E.; Brophy, T. Students Starting University: Exploring Factors That Promote Success for First-Year International and Domestic Students. *Front. Educ.* **2022**, *7*, 779756. [CrossRef]
12. Del Savio, A.A.; Galantini, K.; Pachas, A. Exploring the relationship between mental health-related problems and undergraduate student dropout: A case study within a civil engineering program. *Heliyon* **2022**, *8*, e09504. [CrossRef]
13. Loureiro, A.; Carvalho, C.; Rodrigues, M.d.O. Affirmative Action for Black, Indigenous and Quilombola Students at a Brazilian University. *Societies* **2024**, *14*, 189. [CrossRef]

14. Anastasiou, M. Mapping international students' expectations from the CY hospitality and tourism higher education: An early dropout indicator. *Tour. Hosp. Manag.* **2019**, *25*, 269–290. [CrossRef]
15. da Costa, F.J.; Bispo, M.d.S.; Pereira, R.d.C.d.F. Dropout and retention of undergraduate students in management: A study at a Brazilian Federal University. *RAUSP Manag. J.* **2018**, *53*, 74–85. [CrossRef]
16. Llauró, A.; Fonseca, D.; Romero, S.; Aláez, M.; Lucas, J.T.; Felipe, M.M. Identification and comparison of the main variables affecting early university dropout rates according to knowledge area and institution. *Heliyon* **2023**, *9*, e17435. [CrossRef] [PubMed]
17. Sylvia, S. Seeing Is Believing: Experimental Evidence on the Impact of Eyeglasses on Academic Performance, Aspirations, and Dropout among Junior High School Students in Rural China. *Econ. Dev. Cult. Change* **2019**, *68*, 335–355. [CrossRef]
18. Parviainen, M.; Aunola, K.; Torppa, M.; Poikkeus, A.M.; Vasalampi, K. Symptoms of psychological ill-being and school dropout intentions among upper secondary education students: A person-centered approach. *Learn. Individ. Differ.* **2020**, *80*, 101853. [CrossRef]
19. De Freitas, S.; Gibson, D.; Plessis, C.D.; Halloran, P.; Williams, E.; Ambrose, M.; Dunwell, I.; Arnab, S. Foundations of dynamic learning analytics: Using university student data to increase retention. *Br. J. Educ. Technol.* **2015**, *46*, 1175–1188. [CrossRef]
20. Lindner, C.; Zitzmann, S.; Klusmann, U.; Zimmermann, F. From procrastination to frustration—How delaying tasks can affect study satisfaction and dropout intentions over the course of university studies. *Learn. Individ. Differ.* **2023**, *108*, 102373. [CrossRef]
21. Hla, D.A.; Hindin, D.I. Generative AI & machine learning in surgical education. *Curr. Probl. Surg.* **2025**, *63*, 101701. [CrossRef]
22. Delogu, M.; Lagravinese, R.; Paolini, D.; Resce, G. Predicting dropout from higher education: Evidence from Italy. *Econ. Model.* **2024**, *130*, 106583. [CrossRef]
23. Seo, E.Y.; Yang, J.; Lee, J.E.; So, G. Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. *Heliyon* **2024**, *10*, e30960. [CrossRef] [PubMed]
24. Ye, X.; Zhai, M.; Feng, L.; Xie, A.; Wang, W.; Wu, H. Still want to be a doctor? Medical student dropout in the era of COVID-19. *J. Econ. Behav. Organ.* **2022**, *195*, 122–139. [CrossRef]
25. Shafique, R.; Al-Shamayleh, A.S.; Posa, S.K.; Ishaq, A.; Rustam, F.; Choi, G.S. Advancing ovarian cancer outcomes with CTGAN-enhanced hybrid machine learning approach. *Knowl. Based Syst.* **2025**, *328*, 114206. [CrossRef]
26. Li, D.; Hu, T.; Dong, L.; Ma, S. A generalized generation and evaluation method for cutting process parameter knowledge based on CTGAN. *Robot. Comput. Integr. Manuf.* **2025**, *94*, 102963. [CrossRef]
27. Habibi, O.; Chemmakha, M.; Lazaar, M. Imbalanced tabular data modelization using CTGAN and machine learning to improve IoT Botnet attacks detection. *Eng. Appl. Artif. Intell.* **2023**, *118*, 105669. [CrossRef]
28. Adamecz, A. Longer schooling with grade retention: The effects of increasing the school leaving age on dropping out and labour market success. *Econ. Educ. Rev.* **2023**, *97*, 102487. [CrossRef]
29. Adiputra, I.N.M.; Wanchai, P. CTGAN-ENN: A tabular GAN-based hybrid sampling method for imbalanced and overlapped data in customer churn prediction. *J. Big Data* **2024**, *11*, 121. [CrossRef]
30. Xu, Z.; Xu, Q. Psychological factors influencing successful music learning using deep learning technologies. *Interact. Learn. Environ.* **2024**, *32*, 4652–4662. [CrossRef]
31. Liang, G.; Jiang, C.; Ping, Q.; Jiang, X. Academic performance prediction associated with synchronous online interactive learning behaviors based on the machine learning approach. *Interact. Learn. Environ.* **2024**, *32*, 3092–3107. [CrossRef]
32. Bentegri, H.; Rabehi, M.; Kherfane, S.; Nahool, T.A.; Rabehi, A.; Guermoui, M.; Alhussan, A.A.; Khafaga, D.S.; Eid, M.M.; El-Kenawy, E.S.M. Assessment of compressive strength of eco-concrete reinforced using machine learning tools. *Sci. Rep.* **2025**, *15*, 5017. [CrossRef]
33. Doleck, T.; Agand, P.; Pirrotta, D. Predicting data science performance from log data: Using machine learning. *Educ. Inf. Technol.* **2025**, *30*, 14561–14575. [CrossRef]
34. Wu, X.Y. Exploring the impact of blended collaborative learning on deep learning outcomes: A structural equation modeling approach. *Educ. Inf. Technol.* **2025**, *30*, 13411–13437. [CrossRef]
35. Zahoor, K.; Bawany, N.Z. Explainable artificial intelligence approach towards classifying educational android app reviews using deep learning. *Interact. Learn. Environ.* **2024**, *32*, 5227–5252. [CrossRef]
36. Guleria, P.; Sood, M. Explainable AI and machine learning: Performance evaluation and explainability of classifiers on educational data mining inspired career counselling. *Educ. Inf. Technol.* **2023**, *28*, 1081–1116. [CrossRef] [PubMed]
37. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. *Data* **2022**, *7*, 146. [CrossRef]
38. Han, B.; Feng, G.; Shang, P. Node center identification in complex networks based on Jensen–Shannon divergence. *Commun. Nonlinear Sci. Numer. Simul.* **2025**, *150*, 108977. [CrossRef]
39. de Alencar Ribeiro, A.A.; Fernandes, M.A.; Vedana, K.G.G.; Lira, J.A.C.; Barbosa, N.S.; Rocha, E.P.; Cunha, K.R.F. Mental health and university dropout among nursing students: A cross-sectional study. *Nurse Educ. Today* **2025**, *147*, 106571. [CrossRef]

40. Xing, Y.; Hu, Y.; Zhou, J.Z. Higher education and family background: Which really matters to individual's socioeconomic status development in China. *Int. J. Educ. Dev.* **2021**, *81*, 102334. [CrossRef]
41. Li, J.Y.; Wu, X.X.; Fan, Y.R.; Shi, Y.X. Valuation of the cultural adaptation and psychometric properties of the Chinese version of the hidden curriculum evaluation scale in nursing education. *Nurse Educ. Pract.* **2024**, *75*, 103880. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.