Full research paper

UNPACKING THE BLACK BOX: A HYBRID XAI FRAMEWORK FOR AUTOGLUON-BASED MULTICLASS STUDENT OUTCOME PREDICTION

Marwan Nawae✉
Siripa Chankua
Massaya Longsaman

Faculty of Education and Liberal Arts,
Hatyai University, Thailand

✉ marwan.n@hu.ac.th

ABSTRACT

High student dropout rates remain a significant impediment to achieving the United Nations SDG 4 (equitable education). While Artificial Intelligence (AI) offers robust early risk prediction, the intrinsic black-box nature of high-performing models constrains their transparency. This study designs and investigates a multi-layered Explainable AI (XAI)-based assessment framework to generate actionable insights for student retention. We utilized AutoGluon to construct high-performing multiclass classification models (Graduated, Dropout, or Enrolled) on a higher education dataset. To address the complexity of the AutoGluon-generated models, we employed a hybrid XAI framework that couples global interpretability via a decision tree surrogate model and local interpretability via LIME (Local Interpretable Model-agnostic Explanations). The analysis revealed that models from the Boosting family, particularly XGBoost with bagging level 2, achieved the highest predictive performance (exceeding 0.890 across all metrics). The global analysis demonstrated that academic factors were the primary drivers of prediction, but critical socio-economic factors, such as Tuition fees, also exerted significant influence. Local LIME analysis provided granular, case-specific insights, strongly linking dropout status to first-year academic challenges and to features such as age at enrollment. This integrated XAI approach transforms complex models into an interpretable system, supporting student retention and educational equity (SDG 4).

KEYWORDS

AutoGluon, educational efficiency, Explainable AI (XAI), machine learning, SDG 4, student dropout prediction

HOW TO CITE

Nawae M., Chankua S., Longsaman M. (2026) 'Unpacking the Black Box: A Hybrid XAI Framework for AutoGluon-Based Multiclass Student Outcome Prediction', *Journal on Efficiency and Responsibility in Education and Science*, vol. 19, no. 1, pp. 28–39. http://dx.doi.org/10.7160/eriesj.2026.190103

Article history
Received
October 31, 2025
Received in revised form
March 8, 2026
Accepted
March 9, 2026
Available on-line
March 31, 2026

Highlights

- *A hybrid XAI framework interprets complex AutoML for multiclass prediction.*
- *Boosting model (XGBoost) achieved the highest performance (Accuracy > 0.890).*
- *Global analysis confirms that academic and socio-economic factors are the primary drivers.*
- *Local analysis provides granular, case-specific insights into each class for intervention.*

INTRODUCTION

The pursuit of equitable and inclusive education directly addresses the United Nations Sustainable Development Goal 4 (SDG 4). SDG 4 mandates high-quality, inclusive, and equitable education, promoting lifelong learning opportunities for all. However, achieving this remains challenging due to persistent inequalities in tertiary education. According to the Education at a Glance 2025: OECD Indicators report (OECD, 2025), unequal opportunities significantly hinder the educational attainment of learners from disadvantaged backgrounds. On average across OECD countries, only 26% of young adults whose parents did not complete upper secondary education hold a tertiary qualification. Low completion rates compound this access gap. Newly collected data from over 30 countries show that only 43% of new bachelor's students graduate within the expected duration. Even with three additional years, the completion rate only reaches 70% (OECD, 2025). This high attrition rate disproportionately affects marginalized groups and acts as a major impediment to inclusivity. Beyond academic failure, the socioeconomic consequences are profound. Research indicates that individuals without a degree have a 15% to 20% lower lifetime earning potential than graduates (Krüger et al., 2023). Furthermore, high attrition rates lead to workforce instability and increase the risk of long-term unemployment and intergenerational poverty cycles (Bandala and Andrade, 2017). These economic impacts prove that student retention is not just an academic issue, but a socio-economic necessity.

In this study, educational efficiency is defined as the ability of an institution to use its limited resources, such as time, funding, and teaching staff, to achieve the best possible student outcomes. To ensure high effectiveness, it is necessary to provide an evaluation that offers actionable, interpretable, and timely feedback. This forms the basis for designing better student participation and learning outcomes (Nagy and Molontay, 2024). Despite the growth of educational technology, many current systems still fail to identify at-risk students early enough to help them. Most traditional assessment methods are summative and retrospective, meaning they only look at past performance, often when it is too late to intervene (Ifenthaler and Yau, 2020). There is a clear need for more proactive, efficient tools that provide timely, actionable feedback to keep students motivated (Aulck et al., 2016).

In recent years, Machine Learning (ML) has become a popular tool for predicting student failure. ML models can analyze many factors, such as grades, attendance, and socio-demographic data, to find students who might quit (Realinho et al., 2022; Zanellati et al., 2024). However, most of these models are black boxes. This means they provide a prediction but do not explain why a student is at risk. This lack of transparency creates a trust gap between teachers and students and undermines the moral integrity of AI in assessment, raising concerns about algorithmic bias. Without clear reasons, institutions cannot fulfill their responsibility or decide on the best support, and models might unfairly penalize students based on their background (Arrieta et al., 2020; ElShawi et al., 2021).

To solve this, the field of Explainable Artificial Intelligence (XAI) can be addressed. XAI methods, such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), help turn complex predictions into human-understandable explanations (Padmasiri and Kasthuriarachchi, 2024). By using XAI, educational institutions can improve operational efficiency by focusing limited teaching resources on each student's specific needs rather than a one-size-fits-all approach. XAI provides an enhanced understanding of the complex variables influencing student performance, facilitating early intervention in high-risk cases. Moreover, XAI promotes algorithmic responsibility by allowing educators to detect and mitigate hidden biases (Guevara-Reyes et al., 2025).

This study aims to design and investigate a multi-layered explainable AI-based assessment framework. We are particularly interested in examining how interpretability feedback, derived from this framework, can generate actionable insights to enhance students' learning outcomes, improve retention, and encourage equitable access to quality education.

LITERATURE REVIEW

AI in student performance prediction

Academic achievement and student dropout prediction are important areas in educational data mining. Researchers have used a range of datasets, machine learning models, and explainability techniques to address these issues.

Berens et al. (2019) developed an early detection system (EDS) to identify at-risk students at German universities. Their model, which used administrative data, showed that grades from the first few semesters are strong predictors of a student's likelihood of graduating. However, they warned that these models show patterns but do not explain the exact causes of student dropout. By 2022, Realinho et al. (2022) introduced a large dataset of 4,424 students from a Portuguese school. This dataset includes information about students' families, grades, and the economy. It was designed to help researchers compare different models. Following this, Martins et al. (2023) used this data to make predictions at different times of year. They found that data collected by the end of the first semester gives the most accurate results. They also suggested that models should be adjusted to fit different academic programs. Building on these studies, Villar and de Andrade (2024) compared different algorithms using the same Portuguese data. They found that boosting models, such as LightGBM and CatBoost, worked much better than older methods. These advanced models achieved accuracy scores over 85%, while older methods reached only 70-75%. They also used SHAP to explain which factors were most important. However, they noted that because they used only one dataset, their findings might not apply to every school.

Collectively, these studies show that AI is improving at predicting student success through increasingly complex models. However, a major problem remains: these models are often hard to understand. We need better ways to explain why a specific student is at risk so teachers can provide the right support.

Explainable AI in Classification

In important areas like education, AI models must be accurate and easy to understand. XAI helps people understand how complex black-box models make decisions (Panda and Mahanta, 2023). One common way to do this is by using surrogate models. These are simpler models, such as decision trees, that mimic the logic of complex systems to make them easier for humans to understand (Falvo and Cannataro, 2024). This allows schools to check how the AI works and ensure it is fair. Another popular tool is LIME. LIME explains individual predictions by looking at specific cases (Ribeiro et al., 2016; Falvo and Cannataro, 2024). However, LIME can sometimes produce unstable results due to its data sampling strategy (ElShawi et al., 2021).

In education, combining surrogate models with LIME can provide both a clear overall view and personalized feedback for each student. This helps teachers and students trust the AI system more and ensures it is used ethically.

MATERIALS AND METHODS

This section describes the experimental pipeline, dataset, data preprocessing, model construction, evaluation measures, and the explainability techniques employed in this work, as shown in Figure 1.

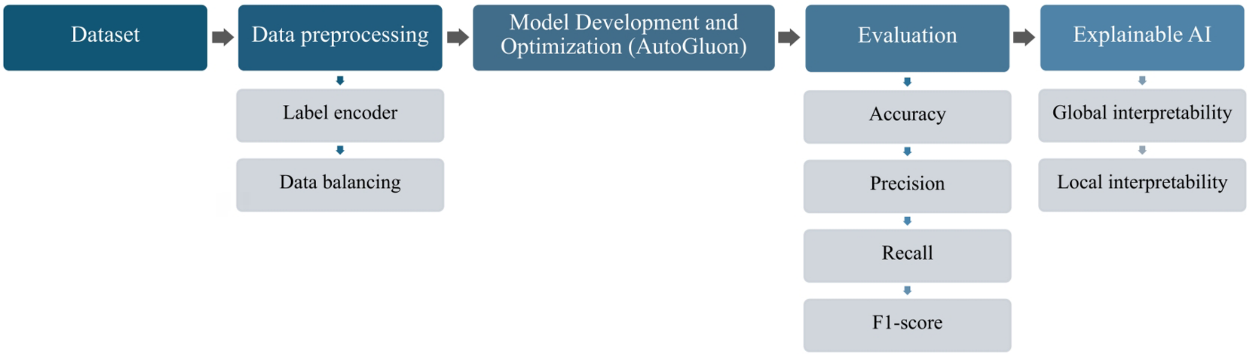


Figure 1: Workflow of the study, 2025 (source: own drawing)

Dataset

In this study, a publication dataset was used as the primary source of data for building machine learning models. The dataset is part of a learning analytics system used at the Polytechnic Institute of Portalegre. It has been used in previous studies to build machine learning models to forecast student academic achievement and dropout (Realinho et al., 2022). Data were acquired for the 2008/2009 to 2018/2019 cohorts. The dataset integrates information from multiple internal and external institutional systems, including the Academic Management System (AMS), Support System for the Teaching Activity (PAE), the General Directorate of Higher Education (DGES), and the Contemporary Portugal Database (PORDATA). The dataset includes 4,424 student records and 35 variables. It covers student demographics, socio-economic factors, and academic performance from the first two semesters. In this study, we use these data to predict three outcomes: Graduate, Dropout, or Enrolled. Table 1 summarizes the key features of the dataset. More specific details about the data characteristics and variables are available in the work of Realinho et al. (2022).

Category	Features
Personal & Demographic Data	Marital status Nationality Gender Age at enrollment International
Socio-Economic & Family Data	Mother’s qualification Father’s qualification Mother’s occupation Father’s occupation Tuition fees up to date Scholarship holder Debtor
Macro-Economic Indicators	Unemployment rate Inflation rate GDP
Academic & Application Factors	Application mode Application order Course Daytime/evening attendance Previous qualification Displaced Educational special needs
Academic Performance Metrics (1st Semester)	Curricular units 1st sem (credited) Curricular units 1st sem (enrolled) Curricular units 1st sem (evaluations) Curricular units 1st sem (approved) Curricular units 1st sem (grade) Curricular units 1st sem (without evaluations)
Academic Performance Metrics (2nd Semester)	Curricular units 2nd sem (credited) Curricular units 2nd sem (enrolled) Curricular units 2nd sem (evaluations) Curricular units 2nd sem (approved) Curricular units 2nd sem (grade) Curricular units 2nd sem (without evaluations)

Table 1: Summary of dataset features and categories, 2025 (source: Realinho et al., 2022)

Data preprocessing

We converted the three student status categorical features, graduate, dropout, and enrolled, into numerical labels (0, 1, and 2, respectively) with one-hot encoding as a preprocessing technique. The encoding is suitable for machine learning algorithms as it preserves the distinct identity of each class without implying an ordinal relationship among them (Panda and Mahanta, 2023).

To address the imbalance in the target class distribution, particularly the significantly lower numbers in the dropout class. We utilized the Synthetic Minority Over-sampling Technique (SMOTE). Previous experience with this dataset indicates that, although SMOTE and ADASYN perform similarly, SMOTE performs slightly better (Villar and de Andrade, 2024). SMOTE generates synthetic minority class examples by interpolating between close existing nearest-neighbor examples in the feature space. The technique is widely known to improve classifier performance, especially on educational datasets. With SMOTE used for class balancing, both classes now have an equal number of instances, as shown in Table 2. Particularly, each class now has 2209 samples, for a total of 6627 instances in the balanced dataset.

Label	Before SMOTE	After SMOTE
Graduate	2209	2209
Dropout	1421	2209
Enrolled	794	2209
Total	4424	6627

Table 2: Data distribution before and after SMOTE balancing, 2025 (source: own data)

Model development

For reliable model selection and optimization, we used AutoGluon, a state-of-the-art automated machine learning (AutoML) tool widely acclaimed for its cutting-edge performance on tabular data (Erickson et al., 2020). By automating the whole modeling pipeline, including diligent preprocessing, smart hyperparameter tuning, and advanced model stacking and ensembling, AutoGluon makes the construction of high-performing classification models both highly scalable and reproducible. Most importantly, this automation enables our study to bypass the tedious, often heuristic process of manual model selection and optimization. With AutoGluon's ability to reliably detect and build strong, often intricate, ensemble models, we ensure that the models analyzed by XAI are empirically the best possible black-box predictors. This provides a solid foundation for interpretability efforts, as the intrinsic complexity of these high-performing AutoML-generated models warrants the use of sophisticated post-hoc XAI tools. Although AutoGluon achieves state-of-the-art predictive accuracy, as an automated black box, it inherently obfuscates decision-making processes; a deficiency we directly overcome with our subsequent multi-XAI framework, which is specifically tailored to make these elaborate models transparent and interpretable.

To assess the predictive capability of the suggested model, we randomly split the dataset into a training set and a hold-out test set, with 80% of the data used for training and 20% for testing. This strategy aligns with best practices for moderately large datasets (about 4,500 instances), where computational efficiency must be traded off against robust performance estimation (Kuhn and Johnson, 2013).

AutoGluon was then fit to the specified 80% training split. In this process, it automatically performed end-to-end hyperparameter tuning, managed complex model ensembling, and conducted internal validation to select and configure the best model architecture. The ultimate, top-performing ensemble model, along with its optimized hyperparameters, was then evaluated only on the remaining, untouched 20% test set to provide an unbiased estimate of its capacity to generalize to new, unseen student data.

Evaluation metrics

To evaluate the model's performance in this multi-class classification, we employed four key metrics, consisting of accuracy, precision, recall, and F1-score, as given in Formulas 1–4 (Birchard et al., 2025). All these measures are based on the following fundamental counts for any class *k*.

True Positives ($TP_{k)}$: Samples assigned correctly as class *k*.

True Negatives (TN_k): Samples assigned correctly as not class *k*.

False Positives (FP_k): Samples that were incorrectly predicted to belong to class *k*.

False Negatives (FN_k): Samples incorrectly predicted as not of class *k*.

The metrics utilized are defined by:

Accuracy: This measure computes the ratio of correctly predicted instances to the total number of predictions, reflecting the overall model performance across all classes.

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN} \qquad (1)$$

Precision: For a particular class, precision calculates the proportion of true positive predictions out of all instances predicted as positive for that class. It indicates the model's precision, or the quality of its positive predictions. In multiclass classification, we typically calculate Precision for each class and then average them.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

Recall: For a particular class, recall measures the proportion of true positive predictions among all actual positive cases. It is a measure of the model's completeness or its ability to include all positive, relevant examples. Similar to Precision, Recall is usually calculated for every class and then averaged.

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

F1-Score: The F1-score is the harmonic mean of precision and recall. It is a balanced measure that accounts for both false positives and false negatives and is especially useful for situations with imbalanced datasets. For multiclass problems, we compute the aggregate F1-score by averaging each class's F1-Score.

$$F1-score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Explainability framework

To address the complexity of machine learning black-box models, this work employs a robust explainability framework that provides both global and local interpretability by coupling a surrogate model with LIME. This framework allows stakeholders not only to understand the general patterns underlying predictions but also the specific reasoning behind each prediction. For this analysis, we selected the top-performing model from each major learning type in the AutoGluon.

Global interpretability

We used a decision tree classifier as a global surrogate model to explain the complex, black-box ensemble model generated by AutoGluon. This is based on the idea of model distillation, in which a simpler, more interpretable model is trained to mimic the decision-making structure of a more complex, opaque model (Guidotti et al., 2019). The first step was to generate a synthetic dataset for training the surrogate model. Specifically, for each instance in our original feature space, we obtained the predicted class label from the trained AutoGluon black-box model. This transformed dataset, which kept the original input features while using the black-box model's predictions as the new target variable, served as the training input for the decision tree.

Then the decision tree was trained to recursively split the feature space, developing a list of interpretable if-then rules. This model allowed the surrogate to capture the complex, often non-linear decision boundaries of the black-box model, clearly showing how interactions among features collectively affected the primary model's outputs. The training objective of this surrogate model was to minimize the variance between its own predictions and those of the original black-box model, typically quantified using a loss function such as cross-entropy. A key consideration in this surrogate modeling method is fidelity, which measures how well the interpretable model mimics the black-box model's predictions (Alangari et al., 2023). In classification problems, fidelity is defined as the surrogate model's prediction accuracy relative to the black-box model's output, not the true labels. High fidelity is essential; low fidelity indicates that the surrogate model does not accurately capture the black-box model's decision-making process, undermining the validity and trustworthiness of the derived explanations. When fidelity is convincingly high, the decision tree surrogate delivers a high-level, human-readable view of the prevailing patterns, feature importance, and interactive effects discovered by the complex ensemble model. This macro-level insight is especially useful for auditing, verifying model behavior, and informing governance in sensitive applications like educational policy (Arjunan, 2021).

Local interpretability

We enhanced the local interpretability of our black-box ensemble model by using LIME. LIME serves as a complementary technique to global explanation techniques by providing local, instance-specific explanations about the model's predictions (Ribeiro et al., 2016). This technique is particularly useful for local explanations, which are necessary for personalized interventions to improve student outcomes.

LIME uses a locally faithful surrogate model to predict a specific outcome. For example, LIME starts by generating a new dataset of perturbed samples around this instance. The perturbations are generated by slight changes to the original feature values, effectively creating synthetic data points that are theoretically close to the instance being analyzed.

For each newly generated perturbed instance, the original black-box model predictions are recorded. This process actually interacts with the black-box model to learn about its behavior in the local neighborhood of the target instance. Next, a simple and interpretable model, typically a linear regression or a shallow decision tree, is trained on these perturbed instances. Notably, each perturbed instance is weighted by its closeness to the original instance, ensuring the surrogate model prioritizes local accuracy. This weighting strategy ensures that the interpretable model captures the behavior of the black-box model correctly in the immediate context of the prediction (Ribeiro et al., 2016). The explanation obtained from this locally trained surrogate conveys feature weights or local rules that explain each input variable's contribution to the prediction of the specific case. This method is highly beneficial when decision-makers, such as university administrators or instructors, require quick, intuitive, and actionable explanations for individual student cases in real-time decision-making scenarios (ElShawi et al., 2021).

RESULTS

This section is structured into two parts. The first part presents the performance evaluation results of the models generated by the AutoGluon framework across various metrics. The second part provides explainability for the highest-performing black-box model of each model type at both global and local levels.

Model performance

As seen in Table 3, almost all models showed strong overall performance, with scores above 0.83 across all metrics. The top three models are XGBoost with bagging level 2 (referred to as XGBoost-B2), Weighted Ensemble with bagging level 3 (referred to as WeightedEnsemble-B3), and Light Gradient Boosting Machine with bagging level, achieving very similar results. They all reached an F1-score of 0.899. The XGBoost-B2 and WeightedEnsemble-B3 models recorded the highest precision (0.901), while all three top models had the same accuracy and recall scores at 0.899.

In the next group, the Neural Network model type also maintained high performance, with scores above 0.80 across all metrics. Neural Network implemented via FastAI with bagging level 2 (referred to as NeuralNetFastAI-B2) achieved high performance, with scores of 0.898, 0.899, 0.898, and 0.898 for accuracy, precision, recall, and the F1-score, respectively.

Conversely, models from the Random Forest and Extra Trees groups showed a decline in performance, with F1-scores falling in the range of 0.875 to 0.880. The Random Forest with bagging level 2 (referred to as RandomForestGini-B2) maintained the highest performance, scoring approximately 0.88 across all metrics. Lastly, the models with the lowest scores are below 0.80 but still above 0.72 for each metric.

Model	Accuracy	Precision	Recall	F1-score
XGBoost (bagging level 2)	0.899	0.901	0.899	0.899
WeightedEnsemble (bagging level 3)	0.899	0.901	0.899	0.899
LightGBM (bagging level 2)	0.899	0.900	0.899	0.899
NeuralNetFastAI (bagging level 2)	0.898	0.899	0.898	0.898
LightGBMXT (bagging level 2)	0.897	0.899	0.897	0.898
LightGBMLarge (bagging level 2)	0.893	0.896	0.893	0.893
CatBoost (bagging level 2)	0.893	0.895	0.893	0.893
RandomForestGini (bagging level 2)	0.881	0.882	0.881	0.880
RandomForestEntr (bagging level 2)	0.880	0.882	0.880	0.879
NeuralNetTorch (bagging level 2)	0.879	0.880	0.879	0.879
ExtraTreesGini (bagging level 2)	0.879	0.880	0.879	0.879
ExtraTreesEntr (bagging level 2)	0.876	0.877	0.876	0.875
LightGBMLarge (bagging level 1)	0.865	0.866	0.865	0.865
WeightedEnsemble (bagging level 2)	0.864	0.866	0.864	0.865
LightGBM (bagging level 1)	0.864	0.864	0.864	0.864
XGBoost (bagging level 1)	0.862	0.863	0.862	0.862
LightGBMXT (bagging level 1)	0.855	0.856	0.855	0.855
CatBoost (bagging level 1)	0.853	0.854	0.853	0.853
ExtraTreesEntr (bagging level 1)	0.850	0.853	0.850	0.850
RandomForestGini (bagging level 1)	0.848	0.851	0.848	0.848
ExtraTreesGini (bagging level 1)	0.847	0.850	0.847	0.848
CatBoost (bagging level 1)	0.845	0.847	0.845	0.845
RandomForestEntr (bagging level 1)	0.844	0.847	0.844	0.845
NeuralNetTorch (bagging level 1)	0.841	0.842	0.841	0.841
NeuralNetFastAI (bagging level 1)	0.833	0.834	0.833	0.834
KNeighborsDist (bagging level 1)	0.777	0.789	0.777	0.776
CatBoost (bagging level 1)	0.725	0.729	0.725	0.724
KneighborsUnif (bagging level 1)	0.722	0.729	0.722	0.722

Table 3: Predictive performance of all AutoGluon models on the test set, 2025 (source: own data)

Explainable AI

Based on the overall performance of all models, we selected the best model from three major learning categories (Boosting-based, Neural Network-based, and Tree-based) to perform the XAI analysis. The selected models are XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2. The kNN-based models were excluded from this interpretability analysis because their performance scores fell below 0.8. The XAI results presented cover both global and local interpretability.

Global interpretability

As shown in Table 4, the decision tree surrogate model indicated that all three black-box models (XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2) exhibited similar fidelity scores: 0.7773, 0.7689, and 0.7755, respectively. This similarity suggests that the complex decision boundaries of the original models can be approximated by a simpler decision tree with comparable accuracy.

Furthermore, the global feature importance analysis derived from the surrogate decision tree model indicated that all three black-box models relied on the same top 10 features. These features, ranked in order of influence, are: curricular units 2nd sem (approved), curricular units 2nd sem (grade), tuition fees up to date, course, curricular units 1st sem (grade), age at enrollment, mother's occupation, unemployment rate, curricular units 2nd sem (evaluations), and GDP. Notably, the feature 'curricular units 2nd sem (approved)' was the most influential factor in the model's decision-making processes, clearly standing out from the other features (see Figures 3 and 4 in the Appendix).

Model	Fidelity
XGBoost (bagging level 2)	0.7773
NeuralNetFastAI (bagging level 2)	0.7689
RandomForestGini (bagging level 2)	0.7755

Table 4: Fidelity scores of surrogate decision trees approximating black-box models, 2025 (source: own data)

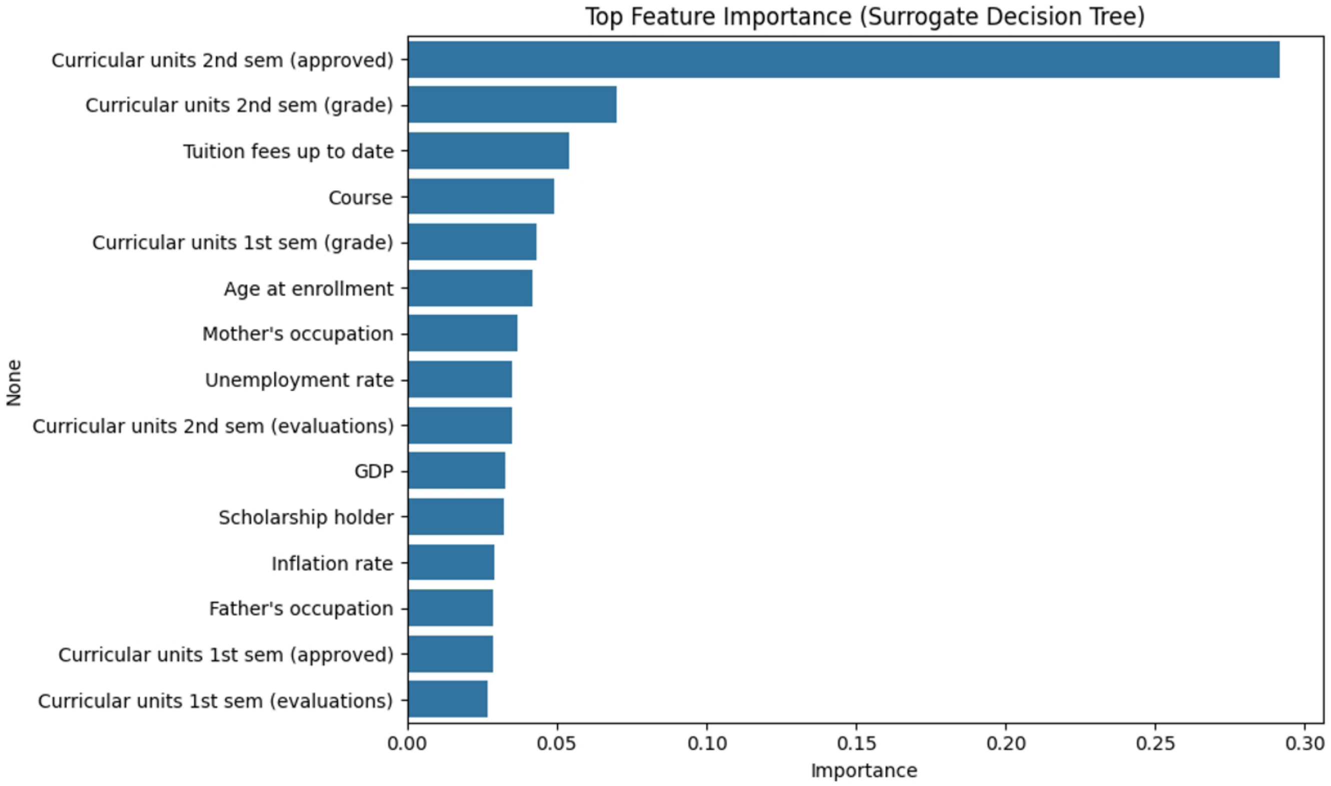


Figure 2: Top feature importance surrogate decision tree, 2025 (source: own data)

Local interpretability

Figure 4 (Appendix) displays the results of the sub-analysis for each classification class using LIME. Unlike the global XAI results, the local explanations showed greater diversity and complexity.

The XGBoost-B2 and NeuralNetFastAI-B2 models consistently achieved high confidence, with prediction probabilities exceeding 0.90 in nearly every instance across all three classes, especially in the graduate and dropout classes. The RandomForestGini-B2 model also performed well but generally showed lower certainty than the other two.

In terms of influential features for prediction, the three models largely shared similar feature sets across classes. The core features common across all three classes included curricular units 2nd sem (approved), curricular units 2nd sem (enrolled), curricular units 2nd sem (grade), curricular units 1st sem (approved), and curricular units 2nd sem (grade). However, specific classes involved additional or unique features: the dropout class also involved the additional features of the mother's qualification and course. In contrast, the enrolled class included education, special needs, and course as additional influential features. These local differences highlight the unique, case-specific logic that each complex model employs when classifying individual data points.

DISCUSSION

Model performance

Based on the experimental results, the predictive performance of almost all models was highly consistent. This consistency is primarily due to the use of Bagging (Bootstrap Aggregating), an ensemble technique in the AutoGluon framework that enhances model stability and accuracy. Bagging works by training multiple base models on random subsets of the data and combining their outputs, effectively reducing prediction variance and preventing overfitting (Sisman et al., 2025).

Specifically, the Gradient Boosting Machines (GBMs) demonstrated superior performance in the dropout prediction task compared to other model categories. The XGBoost-B2 model highlighted the suitability of GBMs for this specific dataset. This finding aligns with the work of Sisman and team, who found that ensemble and boosting models yield strong results for tabular datasets (Sisman et al., 2025). This finding aligns with the core mechanism of the Boosting algorithm, which sequentially builds simple models (weak learners) to iteratively correct the errors of preceding models. This process enables GBMs to capture complex, non-linear relationships in the data (Nguyen and Ngo, 2025).

Concurrently, the NeuralNetFastAI-B2 model achieved performance metrics remarkably close to those of XGBoost-B2. This reflects a growing trend where deep learning architectures designed for tabular data are demonstrating competitive predictive power (Borisov et al., 2024). However, the marginal outperformance of GBMs in this study supports the consensus from several recent benchmark analyses that GBMs remain the state-of-the-art solution for most prediction problems involving tabular data (Grinsztajn et al., 2022).

In contrast, the models in the traditional Tree-based group, which showed reliable performance, exhibited lower overall efficacy when directly compared with the Boosting and Neural Network model groups.

Explainable AI

Global interpretability

When comparing the surrogate model fidelity scores for the three selected black-box models, the values were closely clustered around 0.77. It represents a significant reduction compared to the models' actual classification accuracy achieved by AutoGluon. The substantial gap between the high task accuracy and the lower explanation fidelity highlights the fundamental trade-off between model performance and interpretability. Since black-box models are highly complex, using a simpler decision tree as a surrogate imposes inherent limitations on the surrogate's predictive power. Consequently, as the original model's accuracy increases, its decision structure becomes more intricate, making it increasingly difficult for simple, interpretable models to mimic its behavior faithfully (Awad and Fraihat, 2023).

The analysis of features influencing dropout prediction across all three models revealed that the primary determinants are academic factors, such as Curricular units 2nd sem (approved), Curricular units 2nd sem (grade), Course, and Curricular units 1st sem (grade). This result confirms that academic performance is a central driver of the dropout problem, aligning strongly with previous research in educational data mining (Realinho et al., 2022; Olive et al., 2025).

In addition, socioeconomic factors were identified as highly influential, including Tuition fees up to date and the mother's occupation. The fact that Tuition fees up to date ranked as the third most important feature underscores the significant role of financial status in the model's prediction of student persistence (Olive et al., 2025).

As observed in the decision tree Surrogate model (see Figure 3 in the Appendix), the first partitioning of the data at the root node is defined by Curricular units 2nd sem (approved). Subsequent critical decision paths are governed by Curricular units 2nd sem (grade) and Tuition fees up to date. This hierarchical structure indicates that the model's classification logic systematically combines academic performance and financial status sequentially.

Local Interpretability

For the local interpretability analysis using LIME, we assessed the top features influencing individual predictions for the three best models across three distinct classes: graduate, dropout, and enrolled. While the specific ranking of features exhibits instance-by-instance volatility, the majority of the influential features remain consistent across all models and classes. This section details the key factors driving prediction for each status.

In predicting the graduate status, the models' decisions are predominantly influenced by positive academic performance indicators throughout the first and second semesters. These key factors include: number of curricular units registered, number of curricular units passed in each semester, and grade point average (GPA) for each semester. This result clearly indicates that a strong tendency toward graduation is strongly predicted by sustained academic achievement and consistency across all periods of study, a finding consistent with the previous work, which also identified educational factors as the primary determinant of graduation. (Apumayta et al., 2024; Villar and de Andrade, 2024). Additionally, factors related to age at enrollment within the normal, traditional range for higher education were found to positively influence the prediction of successful completion. Conversely, features such as international status and course type had minimal negative weight in the final prediction.

The prediction of dropout status is driven by a set of academic factors identical to those in the graduate class, but with an opposite directional influence. Specifically, the LIME explanations highlight that academic challenges in the first and second year (for example, failure to pass required credits and enrolled or low semester grades in either the 1st or 2nd semester) are strongly associated with a high likelihood of dropout. Furthermore, the course of study was found to be highly influential, with subjects perceived as complex or requiring a longer study duration significantly increasing the predicted probability of dropout. Intriguingly, socio-economic factors were also prominent, with mothers' occupation and qualifications contributing significantly to dropout prediction, suggesting that family background and economic stability are critical risk indicators (Apumayta et al., 2024). Moreover, age at enrollment outside the typical range for entry into higher education institutions was found to contribute positively to the dropout prediction.

The factors driving the enrolled status share similarities with the dropout factors but exhibit a unique pattern of challenges and resilience. The LIME analysis shows that negative socio-economic factors are influential in these instances, reflecting a context of adversity. However, this is critically counterbalanced by positive academic factors, which remain weighted in a positive direction. This suggests that the enrolled student group comprises individuals with the academic persistence and resilience to continue their studies and achieve academic success despite significant socio-economic barriers (Musaddiq et al., 2022).

Implications for Efficiency and Responsibility

Considering these specific predictors, the analysis presents significant implications for both efficiency and responsibility in education and science. Educational efficiency is directly enhanced because institutions can transition from generalized, resource-intensive support programs to highly targeted early interventions. By focusing on definitively identified academic predictors, such as second-semester grades and approved units, early detection systems optimize the allocation of limited institutional resources. This proactive shift reduces the waste of pedagogical efforts and ensures that support reaches students who need it most, thereby improving overall operational efficiency (Blašková and Staňková, 2023; Nagy and Molontay, 2024). Concurrently, identifying socio-economic vulnerabilities—such as the status of tuition fee payments—underscores deep institutional responsibility. By explicitly recognizing these systemic financial barriers, universities can implement equitable support mechanisms (Ferro and D'Elia, 2020). This ensures that predictive AI is used ethically as a supportive tool for disadvantaged students, rather than a black box that unfairly penalizes them based on their background (ElShawi et al., 2021). This dual approach actively aligns the XAI framework with the pursuit of sustainable, efficient, and responsible education, where data-driven insights serve both institutional performance and social justice.

However, this research faces limitations regarding the surrogate model's performance, which achieved a lower predictive score than the actual performance of the AutoGluon models.

Furthermore, this study is limited to a single open-source dataset from a single higher education institution in Portugal, thereby limiting the generalizability of the findings. Future work should therefore explore more complex surrogate modeling techniques to enhance explanation fidelity and validate the framework's generalizability by applying it to diverse, multi-institutional datasets from varied geographical regions. Crucially, the framework is adaptable for validation using an institution's internal, context-specific dataset to ensure its efficacy and relevance within a specific educational environment.

CONCLUSION

This paper successfully presents the development of student dropout prediction models using the AutoGluon framework. Further, it implements a multi-XAI approach to transform these top-performing black-box models into an interpretable system. The models from the Boosting family generally yielded the best results for the present dataset, with the XGBoost-B2 model achieving the highest overall performance metric for dropout prediction. The XGBoost-B2, NeuralNetFastAI-B2, and RandomForestGini-B2 represent the top-performing models for each learning type, with accuracies of 0.899, 0.898, and 0.881, respectively. This integrated XAI approach confirmed the fundamental trade-off between model accuracy and interpretability. Specifically, the fidelity of the Surrogate Model showed a notable decrease (approximately 0.77) when compared to the actual classification accuracy. The global analysis revealed that academic performance factors were the primary drivers of prediction, yet socio-economic factors, such as Tuition fees, also exerted significant influence. Furthermore, the LIME analysis provided granular, case-specific insights, indicating that both graduate and dropout statuses are strongly linked to academic performance challenges in the first year. Students who earn good grades across both semesters and consistently pass their required units have a high propensity to graduate. Crucially, socioeconomic factors also played a significant role in predicting dropout. Identifying key predictors can further assist in optimizing resource allocation and supporting vulnerable learners. This strategy enhances institutional accountability and aligns the analysis with SDG 4 objectives for global educational equity.

In future work, the prediction models will be refined and tailored to different academic programs within the institution to enhance accuracy and relevance to specific curricula. This targeted approach will enable the creation of actionable, timely feedback to support student retention and contribute to achieving the goal of educational equity (SDG 4).

REFERENCES

Alangari, N., El Bachir Menai, M., Mathkour, H. and Almosallam, I. (2023) 'Exploring Evaluation Methods for Interpretable Machine Learning: A Survey', *Information*, Vol. 14, No. 8, p. 469. https://doi.org/10.3390/info14080469

Apumayta, R. Q., Cayllahua, J. C., Pari, A. C., Choque, V. I., Valverde, J. C. C. and Ataypoma, D. H. (2024) 'University dropout: A systematic review of the main determinant factors (2020–2024)', *F1000Research*, Vol. 13, p. 253. https://doi.org/10.12688/f1000research.154263.2

Arjunan, G. (2021) 'Implementing Explainable AI in Healthcare: Techniques for Interpretable Machine Learning Models in Clinical Decision-Making', *International Journal of Scientific Research and Management (IJSRM)*, Vol. 9, No. 05, pp. 597–603. https://doi.org/10.18535/ijsrm/v9i05.ec03

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, Vol. 58, No. 1, pp. 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016) 'Predicting student dropout in higher education', *arXiv preprint*, arXiv:1606.06364. https://doi.org/10.48550/arXiv.1606.06364

Awad, M. and Fraihat, S. (2023) 'Recursive Feature Elimination with Cross-Validation with Decision Tree: Feature Selection Method for Machine Learning-Based Intrusion Detection Systems', *Journal of Sensor and Actuator Networks*, Vol. 12, No. 5, p. 67. https://doi.org/10.3390/jsan12050067

Bandala, C. A. J. and Andrade, L. A. (2017) 'Education, Poverty and the Trap of Poor Countries in the Face of Development', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 10, No. 4, pp. 101–108. https://doi.org/10.7160/eriesj.2017.100402

Berens, J., Schneider, K., Gortz, S., Oster, S. and Burghoff, J. (2019) 'Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods', *Journal of Educational Data Mining*, Vol. 11, No. 3, pp. 1–41. https://doi.org/10.5281/ZENODO.3594771

Birchard, K., Boccia, C., Lounder, H., Colston-Nepali, L. and Friesen, V. (2025) 'Popfinder: A Highly Effective Artificial Neural Network Package for Genetic Population Assignment', *Molecular Ecology Resources*, Vol. 25, No. 1, p. e14096. https://doi.org/10.1111/1755-0998.14096

Blašková, V. and Staňková, M. (2023) 'Graduate Employability as a Key to the Efficiency of Tertiary Education', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 16, No. 4, pp. 262–274. https://doi.org/10.7160/eriesj.2023.160401

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M. and Kasneci, G. (2024) 'Deep Neural Networks and Tabular Data: A Survey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 6, pp. 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161

ElShawi, R., Sherif, Y., Al-Mallah, M. and Sakr, S. (2021) 'Interpretability in healthcare: A comparative study of local machine learning interpretability techniques', *Computational Intelligence*, Vol. 37, No. 4, pp. 1633–1650. https://doi.org/10.1111/coin.12410

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A. (2020) 'AutoGluon-Tabular: Robust and accurate AutoML for structured data', *arXiv preprint*, arXiv:2003.06505. https://doi.org/10.48550/arXiv.2003.06505

Falvo, F. R. and Cannataro, M. (2024) 'Explainability techniques for artificial intelligence models in medical diagnostic', in: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon: IEEE, pp. 6907–6913. https://doi.org/10.1109/BIBM62325.2024.10821826

Ferro, G. and D'Elia, V. (2020) 'Higher Education Efficiency Frontier Analysis: A Review of Variables to Consider', *Journal on Efficiency and Responsibility in Education and Science*, Vol. 13, No. 3, pp. 140–153. https://doi.org/10.7160/eriesj.2020.130304

Grinsztajn, L., Oyallon, E. and Varoquaux, G. (2022) 'Why do tree-based models still outperform deep learning on typical tabular data?', in: *Advances in Neural Information Processing Systems (NeurIPS 2022)*, Vol. 35, pp. 507–520. https://doi.org/10.48550/arXiv.2207.08815

Guevara-Reyes, R., Ortiz-Garcés, I., Andrade, R., Cox-Riquetti, F. and Villegas-Ch, W. (2025) 'Machine learning models for academic performance prediction: interpretability and application in educational decision-making', *Frontiers in Education*, Vol. 10, p. 1632315. https://doi.org/10.3389/feduc.2025.1632315

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2019) 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys*, Vol. 51, No. 5, pp. 1–42. https://doi.org/10.1145/3236009

Ifenthaler, D. and Yau, J. Y-K. (2020) 'Utilising learning analytics to support study success in higher education: a systematic review', *Educational Technology Research and Development*, Vol. 68, No. 4, pp. 1961–1990. https://doi.org/10.1007/s11423-020-09788-z

Krüger, J. G. C., de Souza Britto Jr, A. and Barddal, J. P. (2023) 'An explainable machine learning approach for student dropout prediction', *Expert Systems with Applications*, Vol. 233, p. 120933. https://doi.org/10.1016/j.eswa.2023.120933

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*, New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3

Martins, M. V., Baptista, L., Machado, J. and Realinho, V. (2023) 'Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education', *Applied Sciences*, Vol. 13, No. 8, p. 4702. https://doi.org/10.3390/app13084702

Musaddiq, M. H., Sarfraz, M. S., Shafi, N., Maqsood, R., Azam, A. and Ahmad, M. (2022) 'Predicting the Impact of Academic Key Factors and Spatial Behaviors on Students' Performance', *Applied Sciences*, Vol. 12, No. 19, p. 10112. https://doi.org/10.3390/app121910112

Nagy, M. and Molontay, R. (2024) 'Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention', *International Journal of Artificial Intelligence in Education*, vol. 34, pp. 274–300. https://doi.org/10.1007/s40593-023-00331-8

Nguyen, N. and Ngo, D. (2025) 'Comparative analysis of boosting algorithms for predicting personal default', *Cogent Economics & Finance*, Vol. 13, No. 1, p. 2465971. https://doi.org/10.1080/23322039.2025.2465971

OECD (2025) *Education at a Glance 2025: OECD Indicators*, Paris: OECD Publishing. https://doi.org/10.1787/1c0d9c79-en

Olive, U., Bosco, M. and Enan, N. (2025) 'Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis', *Journal of Information and Technology*, Vol. 5, No. 4, pp. 31–40. https://doi.org/10.70619/vol5iss4pp31-40

Padmasiri, P. and Kasthuriarachchi, S. (2024) 'Interpretable prediction of student dropout using explainable AI models', in: *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, Colombo: IEEE, pp. 1–7. https://doi.org/10.1109/SCSE61872.2024.10550525

Panda, M. and Mahanta, S. R. (2023) 'Explainable artificial intelligence for healthcare applications using random forest classifier with LIME and SHAP', in: Balas, V. E., Kumar, R. and Srivastava, S. (eds.), *Explainable, Interpretable, and Transparent AI Systems*, Boca Raton: CRC Press, pp. 89–105. https://doi.org/10.1201/9781003442509-6

Realinho, V., Machado, J., Baptista, L. and Martins, M. V. (2022) 'Predicting Student Dropout and Academic Success', *Data*, Vol. 7, No. 11, p. 146. https://doi.org/10.3390/data7110146

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?": Explaining the predictions of any classifier', in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, San Francisco: ACM, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

Sisman, S., Kara, A. and Aydinoglu, A. C. (2025) 'Leveraging spatial data infrastructure for machine learning based building energy performance prediction', *PLOS One*, Vol. 20, No. 1, p. e0335531. https://doi.org/10.1371/journal.pone.0335531

Villar, A. and de Andrade, C. R. V. (2024) 'Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study', *Discovery Artificial Intelligence*, Vol. 4, No. 1, pp. 1–24. https://doi.org/10.1007/s44163-023-00079-z

Zanellati, A., Zingaro, S. P. and Gabbrielli, M. (2024) 'Balancing performance and explainability in academic dropout prediction', *IEEE Transactions on Learning Technologies*, Vol. 17, pp. 2086–2099. https://doi.org/10.1109/TLT.2024.3425959

APPENDIX

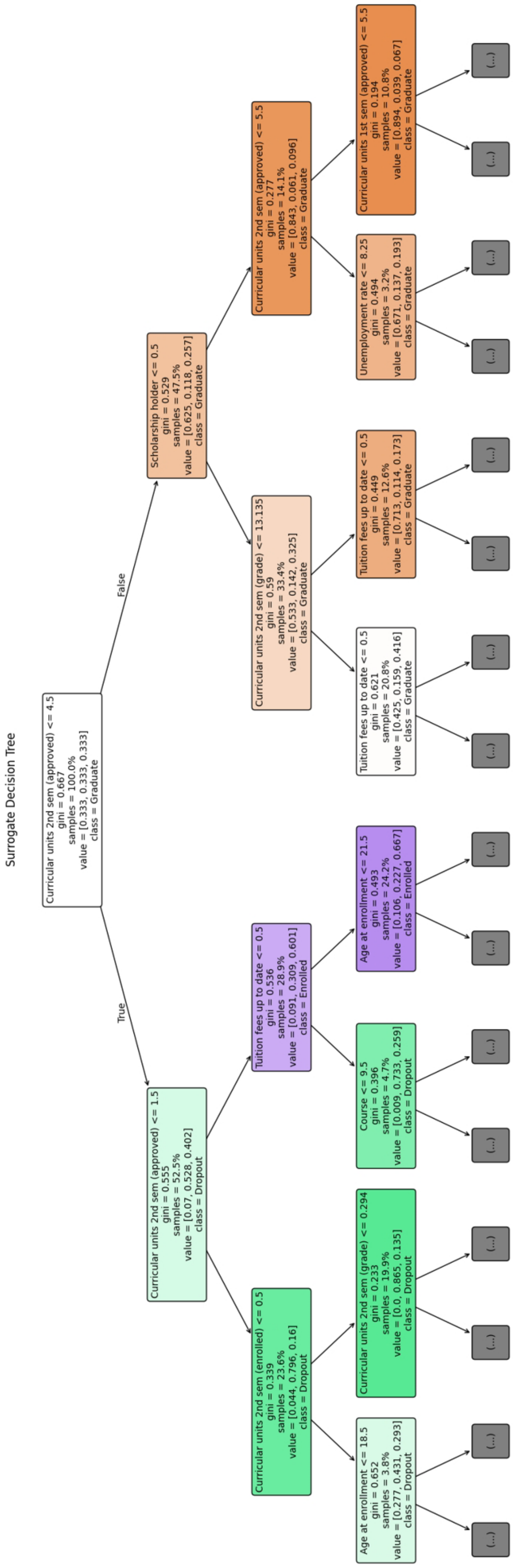
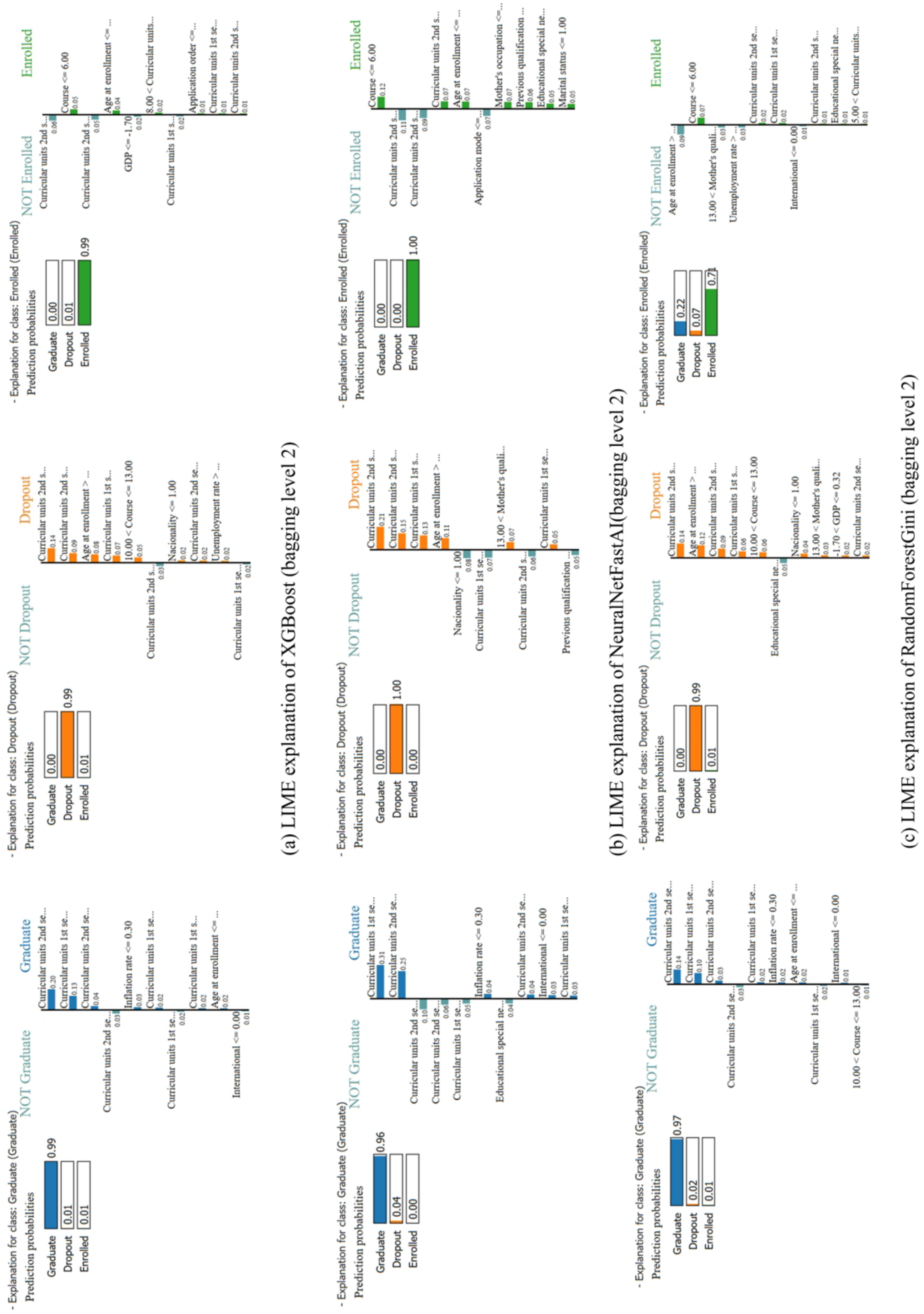

Surrogate Decision Tree
Curricular units 2nd sem (approved) <= 4.5
gini = 0.667
samples = 100.0%
value = [0.333, 0.333, 0.333]
class = Graduate
True
False
Curricular units 2nd sem (approved) <= 1.5
gini = 0.555
samples = 52.5%
value = [0.07, 0.528, 0.402]
class = Dropout
Scholarship holder <= 0.5
gini = 0.529
samples = 47.5%
value = [0.625, 0.118, 0.257]
class = Graduate
Curricular units 2nd sem (enrolled) <= 0.5
gini = 0.339
samples = 23.6%
value = [0.044, 0.796, 0.16]
class = Dropout
Tuition fees up to date <= 0.5
gini = 0.536
samples = 28.9%
value = [0.091, 0.309, 0.601]
class = Enrolled
Curricular units 2nd sem (grade) <= 13.135
gini = 0.59
samples = 33.4%
value = [0.533, 0.142, 0.325]
class = Graduate
Curricular units 2nd sem (approved) <= 5.5
gini = 0.277
samples = 14.1%
value = [0.843, 0.061, 0.096]
class = Graduate
Age at enrollment <= 18.5
gini = 0.652
samples = 3.8%
value = [0.277, 0.431, 0.293]
class = Dropout
Curricular units 2nd sem (grade) <= 0.294
gini = 0.233
samples = 19.9%
value = [0.0, 0.865, 0.135]
class = Dropout
Course <= 9.5
gini = 0.396
samples = 4.7%
value = [0.009, 0.733, 0.259]
class = Dropout
Age at enrollment <= 21.5
gini = 0.493
samples = 24.2%
value = [0.106, 0.227, 0.667]
class = Enrolled
Tuition fees up to date <= 0.5
gini = 0.621
samples = 20.8%
value = [0.425, 0.159, 0.416]
class = Graduate
Tuition fees up to date <= 0.5
gini = 0.449
samples = 12.6%
value = [0.713, 0.114, 0.173]
class = Graduate
Unemployment rate <= 8.25
gini = 0.494
samples = 3.2%
value = [0.671, 0.137, 0.193]
class = Graduate
Curricular units 1st sem (approved) <= 5.5
gini = 0.194
samples = 10.8%
value = [0.894, 0.039, 0.067]
class = Graduate
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Figure 3: Surrogate decision tree plot, 2025 (source: own data)



(a) LIME explanation of XGBoost (bagging level 2)

(b) LIME explanation of NeuralNetFastAI(bagging level 2)

(c) LIME explanation of RandomForestGini (bagging level 2)

Figure 4: LIME explanations in each class for XGBoost, NeuralNetFastAI, and RandomForestGini, 2025 (source: own data)