International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Student Dropout Prediction Using XGBoost and Explainable AI (SHAP): A Case Study of Portuguese Student Dataset

Chandra Kesuma [1*], Vembria Rose Handayani [2]

[1]Department of Computer Technology, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Purwokerto, Indonesia
[2]Department of Information System, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Purwokerto, Indonesia

Authors Scopus ID URL :
Chandra Kesuma[1*], https://www.scopus.com/authid/detail.uri?authorId=58289249700

Email address:
chandra.cka@bsi.ac.id, vembria.vrh@bsi.ac.id

*Corresponding author : chandra.cka@bsi.ac.id

Received: May 10, 2026; **Accepted**: August 21, 2026 ; **Published**: August 24,2026

Abstract: Student dropout is one of the major challenges in higher education, as it can affect students' academic continuity as well as the efficiency of institutional resource management. The availability of academic, administrative, demographic, and socioeconomic data provides opportunities to develop machine learning models capable of identifying students based on their dropout status. However, models with high predictive performance often have limited interpretability, making them difficult to use as an informative basis for higher education decision-makers. This study aims to develop a dropout status classification model using Extreme Gradient Boosting (XGBoost) and to explain the contribution of features to the model's decisions using Explainable AI through the SHapley Additive exPlanations (SHAP) method. The dataset consists of 4,424 students from higher education institutions in Portugal, with 36 predictor variables. The original three-class target, consisting of Dropout, Enrolled, and Graduate, was transformed into a binary classification problem, with Dropout as the positive class and Graduate and Enrolled as the Non-Dropout class. The study compares Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost using accuracy, precision, recall, F1-score, and ROC-AUC. The results of stratified 5-fold cross-validation show that XGBoost achieves the best overall performance, with an accuracy of 88.07%, recall of 77.90%, F1-score of 80.76%, and ROC-AUC of 92.64%. On the testing data, XGBoost achieves an accuracy of 88.14%, precision of 83.03%, recall of 79.23%, F1-score of 81.08%, and ROC-AUC of 93.40%. SHAP analysis indicates that Curricular units 2nd sem (approved) is the feature with the greatest contribution to the model's decisions, followed by Tuition fees up to date, Curricular units 1st sem (approved), Course, and Age at enrollment. Additional experiments removing first- and second-

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

semester academic features resulted in a substantial decrease in model performance. These findings indicate that the model has strong predictive capability but is more appropriately used to predict dropout status based on students' academic trajectories rather than being claimed as an early warning system from the beginning of their studies.

Keywords: Student Dropout, XGBoost, Explainable AI, SHAP, Machine Learning, Educational Data Mining

1. Introduction

Digital transformation in higher education has resulted in an increasing volume and diversity of student data. The data collected by higher education institutions include not only student identities but also demographic characteristics, educational backgrounds, socioeconomic conditions, administrative status, and academic progress. The utilization of such data through machine learning approaches can provide additional information to support academic decision-making processes.

One of the issues that can be analyzed using this approach is student dropout. Dropout is an important concern because students who fail to complete their higher education may be affected themselves, as well as their families and institutions. From the perspective of higher education institutions, a high number of students discontinuing their studies may also be associated with the effectiveness of academic management and institutional resource allocation. Research in Indonesia has shown that analyzing students who are potentially at risk of dropping out can help higher education institutions identify the characteristics of students who require further attention [16], [18].

Machine learning approaches have been widely used to develop student dropout prediction models. Anwar and Permana compared 14 data mining techniques for student dropout prediction and demonstrated the importance of comparing multiple models before selecting an appropriate algorithm [16]. Their study also indicated that ensemble learning and decision tree-based models can provide competitive performance in dropout prediction tasks.

In another study conducted in Indonesia, Putra et al. compared Random Forest and XGBoost using a student dataset consisting of 4,424 records categorized as Dropout, Graduate, and Enrolled. Random Forest achieved higher accuracy than XGBoost under the experimental configuration used in their study [13]. The same dataset was also used by Saputra to analyze student dropout based on early academic performance using Random Forest and Gradient Boosting [15]. The study demonstrated that early academic information has the potential to be used for identifying students at risk of dropout, although classification performance still faced challenges for several classes.
Recent research in Indonesia has also combined XGBoost with Explainable AI. Kurniasari et al. used a dataset of 4,424 students, compared Random Forest and XGBoost, and employed SHAP to identify dominant factors. Their study reported that Random Forest achieved an accuracy of 77.40%, while XGBoost achieved 76.05%, with *Curricular units 2nd sem (approved)* and *Tuition*

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

fees up to date identified as the two dominant factors based on SHAP analysis [14]. These findings indicate that the use of the Portuguese dataset and the combination of XGBoost and SHAP are no longer entirely novel. Therefore, this study is not intended merely to replicate the application of XGBoost and SHAP, but rather to contribute through several aspects, including the comparison of four classification algorithms, XGBoost optimization, evaluation using stratified 5-fold cross-validation, global and local SHAP interpretation, and an additional analysis involving the removal of first- and second-semester features.

Specifically, this study employs Logistic Regression, Random Forest, SVM, and XGBoost within the same experimental framework. This comparison is conducted to provide an empirical basis for determining the primary model. Model comparison is also relevant because algorithmic performance may vary depending on the characteristics of the data and the experimental configuration [16].

XGBoost was selected because it is a gradient boosting method based on decision trees that can handle nonlinear relationships and interactions among variables. Chen and Guestrin introduced XGBoost as a tree boosting system designed to provide efficient and high-performance models for various classification problems [2]. However, model performance alone is insufficient in the context of education. Higher education administrators need to understand why a particular student is classified as being at risk of dropout. Complex models may produce accurate predictions but remain difficult to interpret directly. This condition provides a rationale for incorporating Explainable AI into the modeling process.

SHAP is one of the approaches that can be used to explain the contribution of individual features to model predictions. Lundberg and Lee developed SHAP based on the concept of Shapley values, allowing the contribution of each feature to be analyzed both globally and at the individual level [3]. Arrieta et al. also emphasized interpretability as an important component in the development of trustworthy Artificial Intelligence systems [10].

This study also considers the temporal dimension of prediction. A model incorporating first- and second-semester academic information may achieve high predictive performance; however, it cannot automatically be considered an early warning model from the beginning of a student's college education. Therefore, an additional experiment was conducted by removing first- and second-semester features. This approach was used to determine the extent to which the model depends on academic information that becomes available only after students have undergone part of their academic journey.

Based on these issues, this study aims to:

1. develop a binary classification model to predict Dropout and Non-Dropout status;
2. compare Logistic Regression, Random Forest, SVM, and XGBoost;
3. determine the best-performing XGBoost configuration;
4. evaluate model stability using stratified 5-fold cross-validation;
5. analyze the factors contributing to predictions using SHAP; and

evaluate the limitations of the model for earlier prediction scenarios.

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

2. Literature Review

2.1 Student Dropout Prediction

Student dropout prediction is one of the applications of educational data mining that aims to identify patterns associated with students' academic persistence. The factors used for prediction may be derived from academic, demographic, socioeconomic, and administrative data.

Kemper et al. demonstrated that machine learning can be used to predict student dropout by utilizing student characteristics and academic data [11]. Similarly, Vaarma and Li showed that machine learning models can be applied to predict dropout in higher education and identify risk patterns based on student data [8]. In the Indonesian context, Ramadhani et al. used the C4.5 algorithm to classify students at risk of dropping out. Their study used 976 student records and achieved an accuracy of 98.50%, although the recall for the Dropout class was only 64.44% [18]. This finding demonstrates that high accuracy does not necessarily indicate equally strong performance in identifying dropout students. Therefore, this study employs multiple evaluation metrics, particularly recall and F1-score.

2.2 Portuguese Student Dataset

The dataset used in this study was obtained from the research conducted by Realinho et al., entitled *Predicting Student Dropout and Academic Success*. The dataset consists of 4,424 students, 36 predictor variables, and one target variable. The data include demographic, socioeconomic, macroeconomic, enrollment status, and first- and second-semester academic performance information [1]. The original target variable consists of three categories:

- Dropout
- Enrolled
- Graduate

In this study, the target variable was transformed into a binary classification, with Dropout as the positive class. The Enrolled and Graduate classes were combined into the Non-Dropout class. This transformation was performed to provide a more specific research focus, namely distinguishing students who dropped out from those who were not categorized as Dropout.

2.3 Logistic Regression

Logistic Regression is a classification method used to estimate the probability that an observation belongs to a particular class. For binary classification, the class probability can be expressed as:

[P(Y=1|X)=\frac{1}{1+e^{-z}}]

where:

[z=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n]

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Logistic Regression is used as a comparison model because of its relatively simple mathematical structure, and its results can serve as a baseline for more complex models.

2.4 Random Forest

Random Forest is an ensemble algorithm that constructs multiple decision trees and combines their predictions to produce the final output. Breiman introduced Random Forest as a method that combines multiple tree predictors to obtain more robust predictions [4]. Random Forest has been applied in various student dropout prediction studies. Putra et al. compared Random Forest and XGBoost using the Portuguese student dataset and found that Random Forest achieved higher accuracy under the experimental configuration used in their study [13].

2.5 Support Vector Machine

Support Vector Machine (SVM) is a classification method that determines an optimal hyperplane for separating data classes. Cortes and Vapnik developed the support-vector network as a classification approach capable of operating in high-dimensional feature spaces [5]. SVM is included as the third baseline model to provide a comparison between a margin-based approach and tree-based models.

2.6 XGBoost

XGBoost is a gradient boosting algorithm that builds a predictive model iteratively. Each new tree is designed to improve the errors produced by the preceding model.
In general, the XGBoost model can be expressed as:

[\hat{y}*i*=*sum*{k=1}^{K}f_k(x_i), \quad f_k \in F]

where (K) represents the number of trees and (f_k) represents the (k)-th tree function [2].

The advantages of XGBoost include its ability to handle nonlinear relationships, feature interactions, regularization, and efficient learning. However, the complexity of the model makes individual decisions difficult to explain directly. Sudarman and Budi, in an Indonesian study, applied XGBoost to predict student academic success and achieved an accuracy of 76.8%, demonstrating the potential of XGBoost for academic data analysis [17].

2.7 Explainable AI

Explainable Artificial Intelligence (XAI) is used to improve the transparency of machine learning models. Arrieta et al. explained that XAI aims to provide an understanding of the mechanisms and outputs of AI models so that the resulting decisions can be more easily understood by humans [10].

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

In the educational domain, interpretability is important because model decisions may affect individual students. Therefore, predictions should ideally be accompanied by information regarding the factors contributing to those decisions.

2.8 SHAP

SHAP, or SHapley Additive exPlanations, applies the concept of Shapley values from game theory to determine the contribution of individual features to a model prediction [3].
In simplified form:

[f(x)=\phi_0+\sum_{i=1}^{n}\phi_i]

where:

- (\phi_0) = the model's baseline value;
- (\phi_i) = the contribution of the (i)-th feature to the prediction.

A positive SHAP value indicates that a feature contributes toward the predicted class being analyzed, whereas a negative value indicates a contribution in the opposite direction.
SHAP can be used for two types of interpretation:

1. **Global explanation**, which identifies the most important features across the dataset.
2. **Local explanation**, which explains why a particular individual receives a specific prediction.

A similar approach has been applied in student dropout research in Indonesia. Kurniasari et al. used SHAP to identify dominant features in a student dropout model and found that the number of courses passed in the second semester and tuition fee payment status were among the dominant factors [14].

3. Methods

3.1 Research Design

This study employs a quantitative experimental approach with the following stages:
Dataset → Target Transformation → Preprocessing → Data Splitting → Baseline Models → XGBoost → Hyperparameter Optimization → Cross-Validation → Evaluation → SHAP → Early Prediction Analysis
All experiments were conducted using Python and appropriate machine learning libraries.

3.2 Dataset

The dataset consists of 4,424 students with 36 predictor variables and one target variable.
Table 1. The original target distribution

Target	Number
Graduate	2,209

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Dropout	1,421
Enrolled	794
Total	**4,424**

Table 2. Transformasi Target into a binary classification:

Original Target	Research Target
Dropout	Dropout (1)
Graduate	Non-Dropout (0)
Enrolled	Non-Dropout (0)

Table 3. Binary Classification distribution

Class	Number	Percentage
Dropout	1,421	32.12%
Non-Dropout	3,003	67.88%
Total	**4,424**	**100%**

The class imbalance was taken into consideration when selecting the evaluation metrics.

3.3 Data Splitting

The data were divided into 80% training data and 20% testing data using a stratified train-test split. A random state of 42 was used to ensure the reproducibility of the experiments.
The testing data were not used during the model training process and were therefore reserved for final model evaluation.

3.4 Compared Models

Four classification models were employed:

1. Logistic Regression;
2. Random Forest;
3. Support Vector Machine (SVM); and
4. XGBoost.

The models were compared using the same dataset and validation scheme to ensure a fair comparison of their performance.

3.5 XGBoost Optimization

Table 4. The final XGBoost configuration

Parameter	Value
n_estimators	250
max_depth	4
learning_rate	0.08
subsample	0.90

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

colsample_bytree	0.90
min_child_weight	3
gamma	0.10
scale_pos_weight	1.60
random_state	42

The scale_pos_weight parameter was used to assign greater emphasis to the Dropout class, which has fewer observations than the Non-Dropout class.

3.6 Stratified 5-Fold Cross-Validation

Model validation was performed using 5-fold cross-validation while maintaining the class proportions in each fold. In each iteration, four folds were used for model training and one fold was used for validation. This process was repeated until every fold had served as the validation set. The mean and standard deviation of the evaluation results were calculated to describe the model's performance and stability.

3.7 Evaluation Metrics

The models were evaluated using the following metrics:

Accuracy

[Accuracy=\frac{TP+TN}{TP+TN+FP+FN}]

Precision

[Precision=\frac{TP}{TP+FP}]

Recall

[Recall=\frac{TP}{TP+FN}]

F1-score

[F1=2\times\frac{Precision\times Recall}{Precision+Recall}]

ROC-AUC

ROC-AUC was used to measure the model's ability to distinguish between classes based on predicted probabilities.

The use of multiple evaluation metrics is necessary because accuracy alone may provide an incomplete representation of model performance when the class distribution is imbalanced. Sokolova and Lapalme also emphasized the importance of using multiple performance measures to better understand the characteristics of classification models [6].

3.8 SHAP Analysis

After XGBoost was selected as the primary model, SHAP was employed to analyze the contribution of individual features to model predictions.

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

The analysis was divided into two levels:

Global Explanation

Global feature importance was measured using the mean absolute SHAP value:
[Mean|SHAP_i|=\frac{1}{N}\sum_{j=1}^{N}|\phi_{ij}|]

A higher value indicates a greater contribution of the corresponding feature to the model's predictions.

Local Explanation

SHAP waterfall plots were used to explain individual predictions. The local analysis was performed on:

- one actual Dropout student; and
- one actual Non-Dropout student.

3.9 Early Prediction Analysis

To examine the model's dependence on first- and second-semester academic information, an additional experiment was conducted by removing variables related to academic performance during these semesters.

This experiment was designed to answer the following question:

How does model performance change when first- and second-semester academic information is not yet available?

This analysis is important to prevent the study from making an unsupported claim that the model can predict student dropout from the beginning of a student's enrollment in higher education.

4. Result And Discussion

4.1 Comparison of Four Algorithms

The results of the stratified 5-fold cross-validation are presented in Table 5.

Table 5. Algorithm Performance Comparison

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	87.64%	87.10%	72.27%	78.97%	91.82%
Random Forest	87.64%	87.44%	71.92%	78.90%	92.21%
SVM	87.00%	**89.95%**	67.07%	76.82%	91.32%
XGBoost	**88.07%**	83.86%	**77.90%**	**80.76%**	**92.64%**

XGBoost achieved the highest accuracy, recall, F1-score, and ROC-AUC among the four models.

SVM achieved the highest precision at 89.95%; however, its recall was only 67.07%. In the context of dropout prediction, this condition should be carefully considered because high precision does not necessarily indicate that the model is capable of identifying most students who actually drop out.

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

XGBoost achieved a recall of 77.90%, which was higher than Logistic Regression at 72.27%, Random Forest at 71.92%, and SVM at 67.07%. This finding differs from the study by Putra et al., which reported that Random Forest outperformed XGBoost under their experimental configuration [13]. This difference indicates that algorithmic performance cannot be considered universal and may be influenced by preprocessing, target transformation, parameter settings, validation schemes, and experimental configurations.Therefore, XGBoost was selected not because it is always superior to Random Forest, but because it achieved the best overall performance under the configuration used in this study.

4.2 XGBoost Cross-Validation Results

Table 6. Stratified 5-Fold Cross-Validation Results

Metric	Mean	Std
Accuracy	**88.07%**	**1.44%**
Precision	**83.86%**	**3.05%**
Recall	**77.90%**	**1.62%**
F1-Score	**80.76%**	**2.20%**
ROC-AUC	**92.64%**	**0.76%**

The relatively small standard deviations indicate that the model's performance did not vary substantially across folds. ROC-AUC had a standard deviation of only 0.76%, indicating stable discriminatory performance. These results provide a stronger basis for evaluating model stability than relying solely on a single training-testing split.

4.3 Testing Data Evaluation

Table 7. XGBoost Testing Results

Metric	Value
Accuracy	**88.14%**
Precision	**83.03%**
Recall	**79.23%**
F1-Score	**81.08%**
ROC-AUC	**93.40%**

The model achieved an accuracy of 88.14%, meaning that approximately 88 out of every 100 students in the testing data were correctly classified. The recall of 79.23% indicates that the model was able to identify most students belonging to the Dropout class. The ROC-AUC of 93.40% indicates a high discriminatory capability.

The testing results were also relatively consistent with the cross-validation results, particularly for ROC-AUC, which remained within the range of approximately 92–93%.

4.4 Confusion Matrix

Table 8. XGBoost Confusion Matrix

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Actual / Predicted	Non-Dropout	Dropout
Non-Dropout	**555**	**46**
Dropout	**59**	**225**

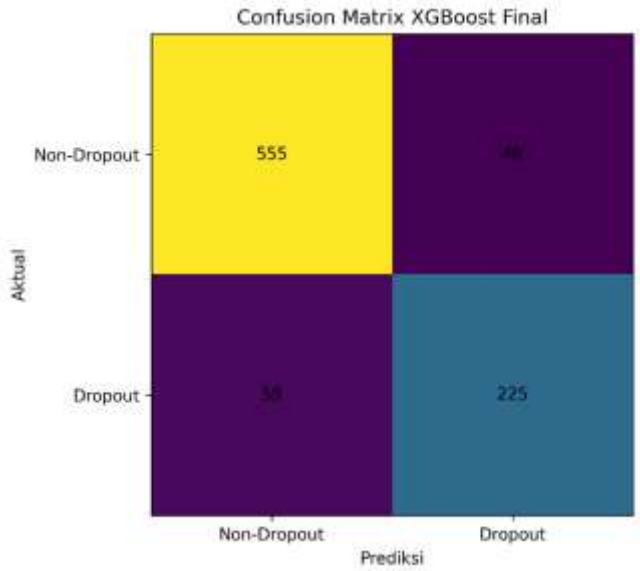

Figure 1. Confusion Matrix XGBoost

Among the 284 Dropout students in the testing data, 225 were correctly identified, while 59 Dropout students were classified as Non-Dropout.
The resulting values were:

- True Positive (TP) = 225
- False Positive (FP) = 46
- True Negative (TN) = 555
- False Negative (FN) = 59

False Negatives are particularly important. If the model is used to support a student monitoring system, these 59 students represent students who actually dropped out but were not identified by the model. Conversely, there were 46 False Positives, representing students predicted as Dropout but actually belonging to the Non-Dropout class. Therefore, the model should be used as a **decision-support tool**, rather than as an automated system that directly determines interventions or actions for students.

4.5 ROC-AUC Analysis

XGBoost achieved a ROC-AUC of 93.40% on the testing data. This result indicates a high ability to distinguish between Dropout and Non-Dropout students.

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

This value was also higher than the mean ROC-AUC obtained through cross-validation, which was 92.64%. Nevertheless, this difference should not be interpreted as evidence that the model has perfect performance. External evaluation using data from different institutions is still required before the model can be generalized to other institutional settings.

4.6 Global SHAP Analysis

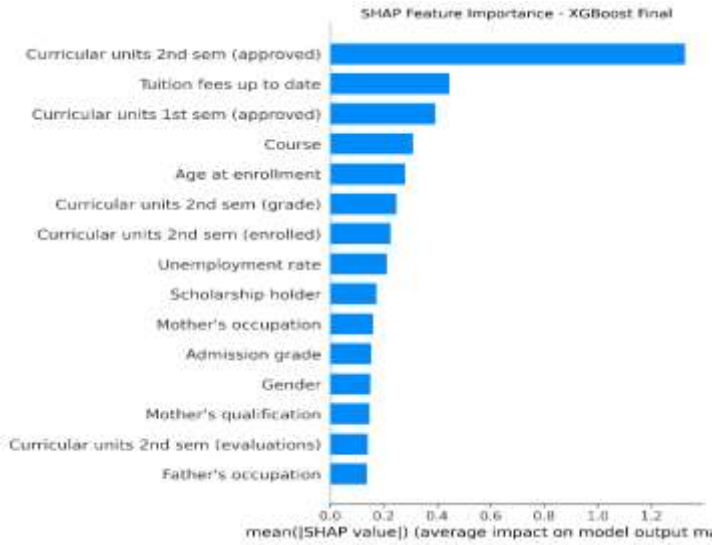

Figure 2. SHAP Analysis

Table 9. Top Ten Features Based on SHAP

Rank	Feature	Mean Absolute SHAP
1	**Curricular units 2nd sem (approved)**	**1.3269**
2	**Tuition fees up to date**	**0.4460**
3	**Curricular units 1st sem (approved)**	**0.3926**
4	**Course**	**0.3105**
5	**Age at enrollment**	**0.2808**
6	**Curricular units 2nd sem (grade)**	**0.2476**
7	**Curricular units 2nd sem (enrolled)**	**0.2271**
8	**Unemployment rate**	**0.2118**
9	**Scholarship holder**	**0.1726**
10	**Mother's occupation**	**0.1617**

The feature *Curricular units 2nd sem (approved)* was the most dominant feature, with a mean absolute SHAP value of 1.3269. This value was substantially higher than the second-ranked feature, *Tuition fees up to date*, which had a value of 0.4460.

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

This result indicates that the number of courses successfully completed during the second semester is highly important to the model when determining student status.
The finding is consistent with Kurniasari et al., who also identified *Curricular units 2nd sem (approved)* and *Tuition fees up to date* as dominant factors using SHAP [14]. However, this study extends that analysis by incorporating a comparison of four algorithms, XGBoost optimization, 5-fold validation, and local SHAP analysis.

4.7 Tuition Payment Status

Tuition fees up to date ranked second, with a mean absolute SHAP value of 0.4460. SHAP direction analysis indicates that an unmet tuition payment status tends to push the prediction toward Dropout, whereas an up-to-date payment status tends to contribute toward Non-Dropout. This finding is noteworthy because it indicates that the model does not rely solely on academic information. Administrative information also contributes to distinguishing the two classes.

However, this result should not be interpreted as evidence that delayed tuition payment directly causes student dropout. SHAP explains the contribution of a feature to the **model's decision**, not a causal relationship.

4.8 Role of First- and Second-Semester Academic Performance

Three academic features were among the seven most important features:

- *Curricular units 2nd sem (approved)*;
- *Curricular units 1st sem (approved)*; and
- *Curricular units 2nd sem (grade)*.

This indicates that academic performance information is an important component of the prediction model.
Saputra's study also demonstrated that early-semester academic performance information can be used in dropout risk modeling [15]. This finding further emphasizes the importance of considering the temporal dimension in student dropout prediction systems.

4.9 Local SHAP Interpretation of a Dropout Student

An actual Dropout student produced a predicted Dropout probability of **90.48%**.
The features providing the largest positive contributions were:

Feature	SHAP Value
Curricular units 2nd sem (approved)	**+1.6935**
Age at enrollment	**+0.8301**
Curricular units 1st sem (approved)	**+0.5815**
Mother's qualification	**+0.2883**
Curricular units 1st sem (enrolled)	**+0.2028**

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Several features contributed in the opposite direction:

Feature	SHAP Value
Course	**−0.6479**
Curricular units 2nd sem (grade)	−0.2202
Tuition fees up to date	−0.2178
Previous qualification (grade)	−0.1919
Curricular units 1st sem (grade)	−0.1697

The interpretation demonstrates that the model does not make its decision based on a single variable. Rather, the prediction results from the aggregation of contributions from multiple features. In this example, *Curricular units 2nd sem (approved)* was the strongest factor pushing the prediction toward Dropout.

4.10 Local SHAP Interpretation of a Non-Dropout Student

For an actual Non-Dropout student, the model produced a Dropout probability of only **7.27%**. The features with the largest negative contributions were:

Feature	SHAP Value
Curricular units 2nd sem (approved)	**−1.4749**
Unemployment rate	−0.3488
Curricular units 1st sem (approved)	−0.2439
Curricular units 2nd sem (grade)	−0.2341
Tuition fees up to date	−0.2050

Conversely, several features contributed toward the Dropout prediction:

- *Course* = +0.4066
- *Admission grade* = +0.3581
- *GDP* = +0.2123
- *Displaced* = +0.1627
- *Curricular units 2nd sem (enrolled)* = +0.0823

The comparison of these two students demonstrates the usefulness of SHAP in explaining predictions at the individual level. Thus, the model output is not limited to a Dropout or Non-Dropout label; it can also be accompanied by information about the factors that contribute most strongly to the model's decision.

4.11 Early Prediction Analysis

An additional experiment was conducted by removing first- and second-semester academic features.

DOI: 10.52362/ijiems.v5i2.2649

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

Table 10. Full Features vs. Without First- and Second-Semester Features

Metric	Full Features	Without Semester 1 & 2
Accuracy	**88.93%**	80.00%
Precision	**86.90%**	76.88%
Recall	**77.11%**	53.87%
F1-Score	**81.72%**	63.35%
ROC-AUC	**93.61%**	84.19%

Removing first- and second-semester academic information resulted in a substantial decrease in model performance.

Accuracy decreased from 88.93% to 80.00%, while recall experienced an even larger decline, from 77.11% to 53.87%. F1-score also decreased from 81.72% to 63.35%. These findings indicate that first- and second-semester academic information is highly important to the model.
This result also has important implications for the use of the term *early prediction*. A model with the highest performance cannot appropriately be described as capable of predicting dropout from the first day a student enters higher education, because several important features only become available after the student has undergone part of the academic process.

Previous research on dropout prediction has also highlighted the importance of defining the prediction time and the availability of information when deploying such models [7], [8], [11]. Therefore, this study is more appropriately described as **student dropout status prediction based on students' academic trajectories**, rather than an *early warning system* from the beginning of the study period.

4.12 Comparison with Previous Studies

Table 11. Comparison with Previous Studies

Study	Dataset/Method	Approach	Main Findings
Anwar & Permana [16]	Student data	14 algorithms	Model comparison is important for determining the appropriate algorithm
Putra et al. [13]	4,424 Portuguese students	RF vs. XGBoost	RF achieved higher performance under their experimental configuration
Saputra [15]	4,424 students	RF & Gradient Boosting	Demonstrated the potential of prediction based on early academic performance
Kurniasari et al. [14]	4,424 Portuguese students	RF, XGBoost, SHAP	RF 77.40%; XGBoost 76.05%; SHAP identified dominant factors
This study	**4,424 Portuguese students**	**LR, RF, SVM, XGBoost + SHAP**	**XGBoost: 88.14% accuracy; 93.40% ROC-AUC**

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

The comparison demonstrates that research results are not necessarily identical even when the same dataset is used.
Differences may arise from preprocessing configurations, target transformations, comparison algorithms, model parameters, validation strategies, and evaluation metrics. Therefore, this study does not claim that XGBoost is universally superior to Random Forest. A more appropriate statement is that **XGBoost achieved the best performance under the experimental configuration used in this study**.

4.13 Research Novelty

The novelty of this study does not lie in the individual use of XGBoost or SHAP, as both approaches have been employed in previous research.
Instead, the contribution of this study lies in the integration of several stages:

1. transforming the original three-class target into a binary classification;
2. comparing four algorithms: Logistic Regression, Random Forest, SVM, and XGBoost;
3. optimizing XGBoost hyperparameters;
4. applying stratified 5-fold cross-validation;
5. evaluating the models using accuracy, precision, recall, F1-score, and ROC-AUC;
6. conducting global SHAP interpretation;
7. conducting local SHAP interpretation for Dropout and Non-Dropout students; and
8. performing an ablation experiment by removing first- and second-semester features to evaluate the limitations of earlier prediction.

Through this approach, the study does not merely address the question, **"How accurately can the model predict student dropout?"** It also addresses:
"Which features contribute most strongly to the prediction?"
and:
"How much does model performance change when first- and second-semester academic information is unavailable?"
This approach provides an important distinction from studies that focus solely on algorithm comparison or accuracy reporting.

5. Conclusions And Suggestions

This study developed a student dropout status prediction model using XGBoost and Explainable Artificial Intelligence (XAI) through SHAP on a dataset of Portuguese higher education students. The dataset consists of 4,424 students with 36 predictor variables. The original target variable, consisting of Dropout, Enrolled, and Graduate, was transformed into a binary classification, with Dropout as the positive class and Graduate and Enrolled as the Non-Dropout class.

A comparison of Logistic Regression, Random Forest, SVM, and XGBoost demonstrated that XGBoost achieved the best overall performance under the experimental configuration used in this

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

study. Based on stratified 5-fold cross-validation, XGBoost achieved an accuracy of 88.07%, recall of 77.90%, F1-score of 80.76%, and ROC-AUC of 92.64%.

On the testing data, XGBoost achieved an accuracy of 88.14%, precision of 83.03%, recall of 79.23%, F1-score of 81.08%, and ROC-AUC of 93.40%. The confusion matrix analysis showed that 225 Dropout students were correctly identified, while 59 Dropout students were still classified as Non-Dropout. This condition indicates that the model has good predictive performance but still presents a False Negative risk that should be considered if the model is used to support student monitoring systems.

SHAP analysis showed that *Curricular units 2nd sem (approved)* was the most dominant feature, followed by *Tuition fees up to date*, *Curricular units 1st sem (approved)*, *Course*, and *Age at enrollment*. These findings indicate that academic performance information and administrative conditions are important components used by the model to distinguish between Dropout and Non-Dropout students. Local SHAP analysis further showed that the model's decision for an individual student results from the combination of contributions from multiple features. Therefore, SHAP can provide additional information regarding the reasons why a particular student receives a Dropout or Non-Dropout prediction.

The experiment conducted without first- and second-semester academic features resulted in a substantial decline in model performance. Recall decreased from 77.11% to 53.87%, while ROC-AUC decreased from 93.61% to 84.19%. These findings indicate that model performance is strongly influenced by academic information that becomes available after students have undergone part of their academic journey.

Therefore, the developed model is more appropriately used as a student dropout status prediction model based on students' academic trajectories rather than being claimed as an *early warning system* from the beginning of the study period.

The model should also not be used as the sole basis for determining actions toward individual students. Machine learning predictions and SHAP interpretations are more appropriately used as supporting information for higher education administrators when identifying students who may require further monitoring or intervention.

Future research could develop models based on multiple prediction stages, such as at enrollment, after the first semester, and after the second semester. Further studies could also conduct external validation using data from higher education institutions in Indonesia, perform statistical significance testing among algorithms, and develop threshold optimization strategies to reduce False Negatives according to institutional requirements.

References

[1] V. Realinho, J. Machado, L. Baptista, and M. V. Martins (2022), “Predicting Student Dropout and Academic Success,” *Data*, vol. 7, no. 11, p. 146, doi: 10.3390/data7110146.

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

[2] T. Chen and C. Guestrin (2016), "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, doi: 10.1145/2939672.2939785.
[3] S. M. Lundberg and S.-I. Lee (2017), "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, pp. 4765–4774.
[4] L. Breiman (2001), "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, doi: 10.1023/A:1010933404324.
[5] C. Cortes and V. Vapnik (1995), "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, doi: 10.1007/BF00994018.
[6] M. Sokolova and G. Lapalme (2009), "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, doi: 10.1016/j.ipm.2009.03.002.
[7] M. V. Martins, L. Baptista, J. Machado, and V. Realinho (2023), "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education," *Applied Sciences*, vol. 13, no. 8, p. 4702, doi: 10.3390/app13084702.
[8] M. Vaarma and H. Li (2024), "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technology in Society*, vol. 76, p. 102474, doi: 10.1016/j.techsoc.2024.102474.
[9] M. Nagy and R. Molontay (2024), "Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 274–300, doi: 10.1007/s40593-023-00331-8.
[10] A. B. Arrieta et al. (2020), "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, doi: 10.1016/j.inffus.2019.12.012.
[11] L. Kemper, G. Vorhoff, and B. U. Wigger (2020), "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, doi: 10.1080/21568235.2020.1718520.
[12] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal (2023), "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 233, p. 120933, doi: 10.1016/j.eswa.2023.120933.
[13] L. G. R. Putra, D. D. Prasetya, and M. Mayadi (2025), "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 147–157, doi: 10.29407/intensif.v9i1.21191.
[14] N. D. Kurniasari, T. B. Rohman, A. Widianto, A. Shobikah, and G. Triono (2026), "Analisis Faktor Dominan Prediksi Dropout Mahasiswa Menggunakan Random Forest, XGBoost dan XAI," *Jurnal Sains Informatika Terapan*, vol. 5, no. 2, doi: 10.62357/jsit.v5i2.1183.
[15] P. S. Saputra (2026), "Analisis Prediktif Dropout Mahasiswa Berdasarkan Kinerja Akademik Semester Awal Menggunakan Machine Learning," *Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI)*, vol. 7, no. 1, doi: 10.30998/jrami.v7i01.791.

International Journal of Informatics, Economics, Management and Science
http://journal.stmikjayakarta.ac.id/index.php/ijiems
E-ISSN: 2809-8471 (online), **P-ISSN:** 2809-9281 (Print)
DOI: 10.52362/ijiems.v5i2.2649
Volume 5 , Issue 2 , August 2026, pp. 180-198

[16] M. T. Anwar and D. R. A. Permana (2021), "Perbandingan Performa Model Data Mining untuk Prediksi Dropout Mahasiswa," *Jurnal Teknologi dan Manajemen*, vol. 19, no. 2, pp. 87–94, doi: 10.52330/jtm.v19i2.34.

[17] E. J. Sudarman and S. Budi (2023), "Pengembangan Model Kecerdasan Mesin Extreme Gradient Boosting untuk Prediksi Keberhasilan Studi Mahasiswa," *Jurnal STRATEGI*, vol. 5, no. 2, pp. 297–314.

[18] A. Ramadhani, R. F. Noor, D. Vernanda, and T. Herdiawan (2024), "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma C4.5 di Politeknik Negeri Subang," *Jurnal Tekno Kompak*, vol. 18, no. 1, pp. 101–112, doi: 10.33365/jtk.v18i1.3439.