Nghiên cứu xây dựng mô hình học máy dự báo khả năng sinh viên bỏ học và thực nghiệm tại Trường Đại học Quy Nhơn

Sinh Phan Đình[1,*], Trần Hoàng Việt[1]

[1]Khoa Công nghệ Thông tin, Trường Đại học Quy Nhơn

[*]Email: phandinhsinh@qnu.edu.vn

Ngày nhận bài: dd/mm/yyyy; Ngày sửa bài: dd/mm/yyyy;

Ngày nhận đăng: dd/mm/yyyy; Ngày xuất bản: dd/mm/yyyy

TÓM TẮT

Nghiên cứu này phát triển mô hình học máy Extreme Gradient Boosting (XGBoost) để dự báo khả năng sinh viên bỏ học sớm tại Trường Đại học Quy Nhơn. Mục tiêu là xây dựng và đánh giá mô hình dựa trên dữ liệu thực tế từ hệ thống quản lý đào tạo của trường, đồng thời so sánh hiệu quả của XGBoost với các mô hình khác như Random Forest và Logistic Regression. Phương pháp nghiên cứu sử dụng dữ liệu từ 7.523 sinh viên, qua các bước tiền xử lý như loại bỏ giá trị ngoại lai và áp dụng kỹ thuật Synthetic Minority Over-sampling Technique (SMOTE) để giải quyết vấn đề mất cân bằng lớp. Kết quả cho thấy mô hình XGBoost đạt hiệu suất vượt trội với các chỉ số Precision, Recall và AUC-ROC gần như hoàn hảo. Mô hình này không chỉ giúp dự báo nguy cơ bỏ học chính xác mà còn có thể ứng dụng vào việc hỗ trợ sinh viên sớm, giúp nâng cao chất lượng giáo dục và quản lý đào tạo tại trường.

Từ khóa: *XGBoost, Dự báo bỏ học, Machine Learning, Random Forest, Logistic Regression*.

A Study on Developing a Machine Learning Model to Predict Student Dropout Risk: An Empirical Investigation at Quy Nhon University

Sinh Phan Đình[1,*], Trần Hoàng Việt[1]

[1]*Faculty of Information Technology, Quy Nhon University*

[*]*Email: phandinhsinh@qnu.edu.vn*

Received: dd/mm/yyyy; Revised: dd/mm/yyyy;

Accepted: dd/mm/yyyy; Published: dd/mm/yyyy

ABSTRACT

This study develops an Extreme Gradient Boosting (XGBoost)–based machine learning model for predicting early student dropout at Quy Nhon University. The objective is to construct and evaluate the proposed model using real-world data obtained from the university's academic management system, and to compare its performance with benchmark models, including Random Forest and Logistic Regression. The study utilizes a dataset of 7,523 students and applies preprocessing techniques such as outlier removal and the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. Experimental results indicate that the XGBoost model consistently outperforms the comparative models, achieving near-perfect Precision, Recall, F1-score, and AUC–ROC values. The proposed model not only enables accurate prediction of student dropout risk but also supports early intervention strategies, thereby contributing to improved educational quality and more effective data-driven academic management.

Keywords: *XGBoost, Dropout Prediction, Machine Learning, Random Forest, Logistic Regression*.

1. INTRODUCTION

Student dropout is one of the most critical issues in higher education, leading to inefficient use of resources, reduced graduation rates, and negative impacts on the reputation of educational institutions. Globally, university dropout rates range from 20% to 50%, depending on factors such as country, field of study, and educational conditions[1]. In Vietnam, this phenomenon has shown an increasing trend, particularly in institutions with large enrollment scales or diverse student populations from different regions. According to statistics from the Ministry of Education and Training, the dropout rate in 2024 was approximately 18%.

Early prediction of students at risk of dropout enables universities to timely implement appropriate interventions, such as academic support, psychological counseling, or financial assistance. However, accurately predicting dropout behavior remains a challenging problem, as student data are typically high-dimensional and nonlinear, and are influenced by multiple factors, including academic performance, family background, learning motivation, and social conditions.

In this context, advances in data science and machine learning have introduced new methodological approaches. Algorithms such as Random Forest, Logistic Regression, Support Vector Machine, and particularly XGBoost have demonstrated strong capabilities in handling complex data and predicting educational outcomes. However, studies applying XGBoost to student dropout prediction in the Vietnamese context remain limited.

The objective of this study is to develop and evaluate an XGBoost-based machine learning model for predicting early student dropout at Quy Nhon University. In addition, the performance of the proposed XGBoost model is compared with two baseline models, namely Random Forest and Logistic Regression, to demonstrate its effectiveness and superiority.

2. METHOD DESCRIPTION

2.1. Research Approach

This study employs a quantitative research methodology, utilizing real-world data extracted

from the Quy Nhon University academic management system. The research process involves key stages such as data collection, preprocessing, model development, training, evaluation, and result interpretation.

The study conducts experiments on three machine learning models: Logistic Regression[2], Random Forest[3,4], and XGBoost[5,6]. To evaluate the predictive performance of these models, the research employs multiple metrics, including the confusion matrix, precision, recall, the AUC-ROC curve, and the F1 score[7,8].

The confusion matrix is used to summarize the classification performance of a machine learning model by comparing predicted labels with the ground-truth values in the test dataset.

In a real-world dataset containing two classes, these are typically labeled as the positive class and the negative class. The classification results predicted by the model on the test dataset are likewise divided into the same two labeled classes (Table 1).

Table 1. Confusion matrix for a dataset with two labeled classes

Actual\Predicted	Positive class	Negative class
Positive class	TP	FN
Negative class	FP	TN

In this context, TP (True Positive) represents the total number of cases where both the actual and predicted outcomes correctly correspond to the positive class; TN (True Negative) denotes the total number of cases where both the actual and predicted outcomes correctly correspond to the negative class; FP (False Positive) refers to the total number of cases in which observations belonging to the negative class are incorrectly predicted as positive; and FN (False Negative) represents the total number of cases in which observations belonging to the positive class are incorrectly predicted as negative. Precision (the proportion of correctly predicted positive cases) measures, among all the instances predicted as positive, how many are truly positive, and is calculated using the following formula.

Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (True Positive Rate) measures, among all actual positive cases, how many were correctly predicted, and is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Accuracy represents the overall correctness of the model and is computed as:

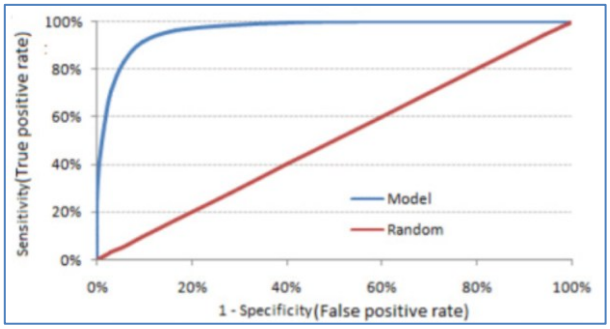
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

F1-score provides a harmonic balance between Recall and Precision, serving as a comprehensive indicator for selecting the most effective model. The higher the F1-score, the better the model's performance. It is calculated as follows:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{2 \times \text{Precision} + \text{Recall}}$$

Receiver Operating Characteristic (ROC) curve: The ROC curve illustrates the trade-off between the true positive rate and the false positive rate as the decision threshold varies. Area Under the ROC Curve (AUC-ROC): The AUC-ROC represents the area under the ROC curve; the larger this area, the better the selected model (Figure 1).

Figure 1. ROC-Receiver Operating Characteristic[8]



To enhance the transparency and interpretability of the nonlinear XGBoost model, this study employs the SHAP (SHapley Additive exPlanations) method for model explanation. All machine learning models used in this research are implemented in the Python programming language for data analysis.

2.2. Dataset and Data Preprocessing

The dataset comprises 7,523 full-time undergraduate students at Quy Nhon University from the admission cohorts spanning 2020 to 2021, extracted from the university's academic management system to ensure that current learning trends are accurately represented (Table 2).

Table 2. Description of information and data types of the attributes

Variable Groups	Attributes	Data types	Description
Personal Information	StudentID	Categorical	Each student ID is unique
	Gender	Categorical	Gender of the student: Male (or Female)
	Ethnicity	Categorical	Ethnicity: Kinh (or other ethnicities)
	Religion	Categorical	Religion: None (or religious)
	Region	Categorical	Enrollment region: (1, 2, 2NT)
	EnrollmentYear	Categorical	Year of student enrollment
	IndustryCode	Categorical	Program code for the student's academic participation
	Aspiration	Categorical	Student's aspirations (NV1, NV2, …)
Admission Information	EntranceScore_1-EntranceScore_3	Numerical	Scores for each subject in the admission evaluation
	SumScore	Numerical	Total score for the admission process
Academic Performance	GPA4_1-GPA4_4	Numerical	GPA for the first to fourth semesters
	Rating_1- Rating_4	Categorical	Rating for each semester ("Excellent" to "Poor")
Academic Credits	CreditsRegistered_1-CreditsRegistered_4	Numerical	Number of credits registered per semester
	CreditsEarned_1-CreditsEarned_4	Numerical	Cumulative credits per semester
Academic Warning	TermStatus_1-TermStatus_4	Numerical	Academic warning for each semester (1 to 4): 1 = Warning, 0 = No warning
Target Variable	Drop	Numerical	1 = Dropout, 0 = Continue Studying

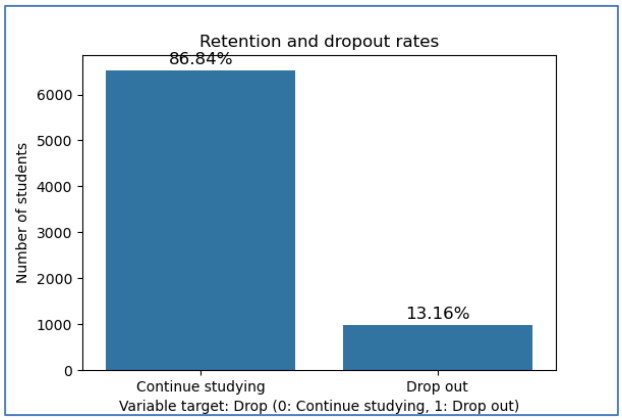
To prepare the dataset for analysis, data cleaning was performed by removing duplicate records and outliers. Categorical variables were encoded using label encoding to convert them into numerical representations compatible with algorithms such as XGBoost. Missing values in score- and credit-related attributes were imputed using mean values, thereby minimizing information loss without introducing significant bias into the data distribution. All numerical features were normalized using the Min–Max scaling method to rescale values to the range [0, 1], which facilitates faster convergence of the machine learning models and improves overall performance.

Additional academic warning features were defined for each semester. Specifically, the academic warning indicator for the first semester (TermStatus_1) was set to 1 if GPA4_1 < 0.8, while the indicators for subsequent semesters (TermStatus_i) were set to 1 if GPA4_i < 1.0 (i = 2–4), in accordance with undergraduate academic regulations.

The target variable for predicting whether a student drops out or continues enrollment was defined by introducing the data field *Drop* into the dataset. A student was labeled as dropout (*Drop* = 1) if at least one semester academic warning indicator equaled 1; otherwise, the student was labeled as non-dropout (*Drop* = 0).

The original dataset was partitioned into training and testing sets using an 80:20 split. Within the dataset, 86.84% of the instances were labeled as *Continue Studying*, while 13.16% were labeled as *Dropout* (Figure 2).

Figure 2. Label Distribution in the Dataset



The class distribution between *Continue Studying* and *Dropout* exhibits a pronounced class imbalance. Therefore, the SMOTE[9] was applied during data analysis to generate synthetic samples for the *Dropout* class, thereby addressing the data imbalance issue prior to model training.

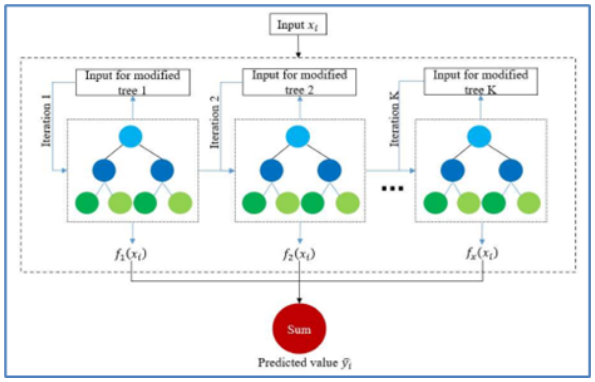
2.3. Machine Learning Models and Techniques

In this study, three machine learning models were implemented within a unified pipeline to predict early student dropout.

Specifically, XGBoost is an ensemble learning algorithm developed based on the Gradient Boosting Decision Tree framework. Introduced by Chen and Guestrin in 2016, XGBoost aims to optimize computational efficiency and generalization performance, and has become one of the most powerful algorithms for classification and regression tasks[5].

XGBoost operates under the boosting paradigm, in which multiple decision trees are constructed sequentially. Each subsequent tree focuses on reducing the errors of the preceding trees by optimizing a loss function using gradient descent (Figure 3).

Figure 3. XGBoost algorithm diagram



The overall learning model can be expressed as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F}.$$

$$\mathcal{F} = \{f(x) = w_{q(x)}\}, q: \mathbb{R}^m \to T, w \in \mathbb{R}^T.$$

Where

- $\hat{y}_i$ denotes the predicted value for the i-th instance;
- m is the number of features;
- $D = \{(x_i, y_i)\}$ represents the training dataset with $| D | = n$, $x_i \in \mathbb{R}^m$, and $y_i \in \mathbb{R}$;
- f_k is the k-th decision tree;
- T denotes the number of leaf nodes in a tree;
- w_i is the weight associated with the i-th leaf node;
- q represents the tree structure that maps an input instance to a corresponding leaf node;
- $\mathcal{F}$ denotes the space of all possible trees.

The learning objective function consists of two components:

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where

$$\Omega(f_k) = \gamma T + \frac{1}{2} \| w \|^2.$$

In this formulation,

- $l(y_i, \hat{y}_i)$ denotes the loss function;
- n is the number of training samples;
- $\hat{y}_i$ and y_i represent the predicted and ground-truth values, respectively;
- $\Omega(f_k)$ is the regularization term used to control the complexity of the k-th tree.

The model incorporates both L1 (Lasso) and L2 (Ridge) regularization terms into the objective function to control model complexity and mitigate overfitting, particularly in scenarios involving high-dimensional data or complex nonlinear relationships. This regularization strategy contributes to reducing variance while also alleviating bias, thereby optimizing the overall predictive performance of the model.

To achieve optimal performance with XGBoost, hyperparameter tuning of *max_depth*, *learning_rate*, *n_estimators*, and *subsample* is essential.

Among these parameters, *max_depth* specifies the maximum depth of each decision tree. Larger

values allow the model to capture complex nonlinear relationships but may increase the risk of overfitting, whereas smaller values reduce model complexity at the expense of potential underfitting. In practice, *max_depth* is typically set within the range of 3 to 10.

The *learning_rate* parameter controls the contribution of each newly added tree during the boosting process. Smaller values (0.01–0.1) lead to slower but more stable learning and typically require a larger number of trees, whereas larger values (>0.3) accelerate learning but increase the risk of overfitting. Consequently, this parameter is commonly tuned in conjunction with *n_estimators*.

The *n_estimators* parameter specifies the number of trees to be trained. A larger number of trees can improve predictive accuracy but increases computational cost and may also elevate the risk of overfitting. Accordingly, when the *learning_rate* is reduced, *n_estimators* is typically increased to maintain model performance.

The *subsample* parameter specifies the fraction of training samples randomly selected for each tree. Values less than 1.0 help reduce overfitting by introducing diversity among trees; however, excessively low values (e.g., below 0.5) may result in the loss of important information. In practice, *subsample* is commonly set within the range of 0.6 to 0.9.

These hyperparameters are typically tuned using grid search in conjunction with cross-validation, with evaluation based on metrics such as AUC and F1-score to ensure an appropriate balance between predictive accuracy and generalization capability.

The Random Forest (RF) machine learning model, proposed by Breiman in 2001, is based on the principle of combining multiple weak learners to form a strong learner[3]. Random Forest constructs an ensemble of decision trees and aggregates their predictions through majority voting for classification tasks or averaging for regression tasks.

Random Forest is constructed based on two key techniques:

Bagging (Bootstrap Aggregating): From the original dataset, the algorithm generates multiple subsets through bootstrap sampling with replacement. Each subset is used to train an independent decision tree.

Random Feature Selection: At each node of a tree, only a random subset of features is considered to determine the optimal split. This strategy reduces correlation among trees and enhances the diversity of the ensemble.

The final prediction is determined as:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_k(x)\}$$

where $h_i(x)$denotes the prediction of the i-th decision tree. Random Forest typically achieves higher accuracy than a single decision tree by effectively reducing overfitting.

The algorithm performs well on high-dimensional data and is capable of handling both categorical and numerical variables.

Logistic Regression (LR)[2] models the relationship between a binary dependent variable $y$and a set of independent variables $X = (x_1, x_2, \dots, x_n)$ using the sigmoid function:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

where $\beta_0, \beta_1, \dots, \beta_n$ are the parameters to be estimated. The sigmoid function constrains the output to the interval $(0,1)$, making it suitable for probabilistic interpretation.

The model parameters are estimated using the maximum likelihood estimation method, which seeks the parameter set that maximizes the probability of the observed data. This optimization process is typically carried out using iterative algorithms such as Gradient Descent or the Newton–Raphson method.

The models were trained using a 60/20/20 data split for training, validation, and testing, respectively, with stratified sampling to preserve label distributions across subsets. To ensure result stability, 5-fold cross-validation was employed throughout the training process. Performance metrics, including Accuracy, Precision, Recall, and F1-score, were computed to evaluate model effectiveness.

3. METHOD EVALUATION

3.1. Model Comparison Results

The performance evaluation results of the three models—Logistic Regression, Random Forest, and XGBoost—indicate that XGBoost consistently outperforms the others in distinguishing between the *Dropout* and *Continue Studying* classes. Specifically, XGBoost achieves an accuracy of 0.996, demonstrating its high classification capability on the test dataset. The model's Precision (0.995) and Recall (0.997) further indicate a well-balanced capability in identifying both positive and negative classes, while effectively minimizing the number of

misclassifications, including false positives and false negatives. The F1-score of XGBoost reaches 0.996, indicating a strong balance between Precision and Recall, which is particularly important in imbalanced classification problems such as student dropout prediction.

Table 3. Performance Results of Prediction Models

Models	Accuracy	Precision	Recall	F1
LR	0.978	0.987	0.969	0.978
RF	0.987	0.987	0.988	0.987
XGBoost	0.996	0.995	0.997	0.996

The performance metrics obtained for XGBoost demonstrate its strong discriminative capability between the two classes, achieving very high effectiveness in correctly identifying students at risk of dropout.

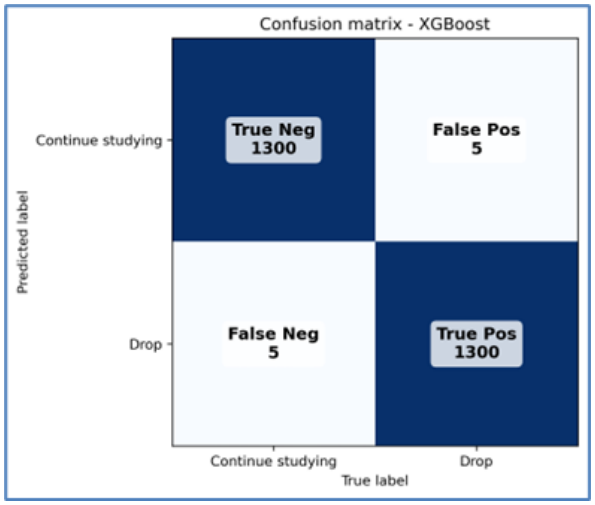
Although Random Forest also exhibits high performance, XGBoost remains superior due to its boosting mechanism and robust optimization strategy, which collectively enhance classification accuracy.

3.2. Confusion Matrix Analysis

The confusion matrix results for the three models—XGBoost, Random Forest, and Logistic Regression—demonstrate their effectiveness in classifying students at risk of dropout.

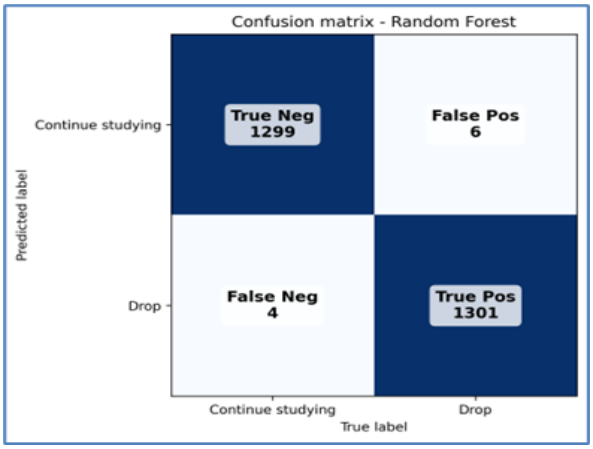
XGBoost exhibits superior performance, with a high number of true positives (TP = 1300) and very low false positives (FP = 5), along with only five false negatives (FN = 5) (Figure 4). These results indicate that the model achieves highly accurate classification with minimal errors.

Figure 4. Confusion Matrix of the XGBoost



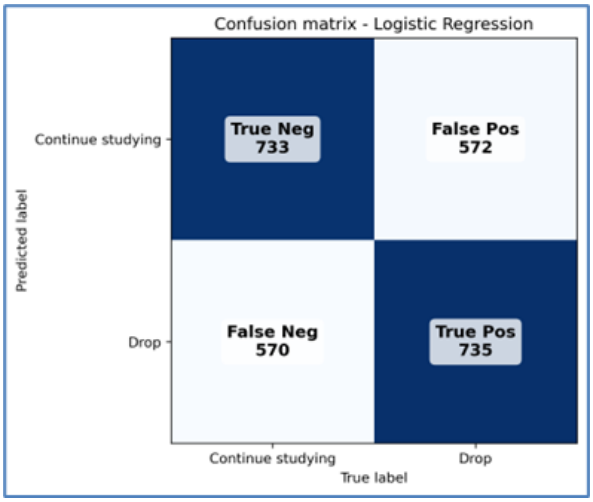
Random Forest also yields favorable results, with TP = 1300 and FP = 6; however, its performance is slightly inferior to that of XGBoost, although the difference is marginal (Figure 5).

Figure 5. Confusion Matrix of the Random Forest



Logistic Regression exhibits substantially inferior performance, with only 733 true positives, while both false positives and false negatives are considerably high (572 and 570, respectively), indicating a high misclassification rate and consequently reduced overall accuracy. This result suggests that Logistic Regression is less effective in distinguishing between students at risk of dropout and those who continue their studies compared with the other two models (Figure 6).

Figure 6. Confusion Matrix of the Logistic Regression



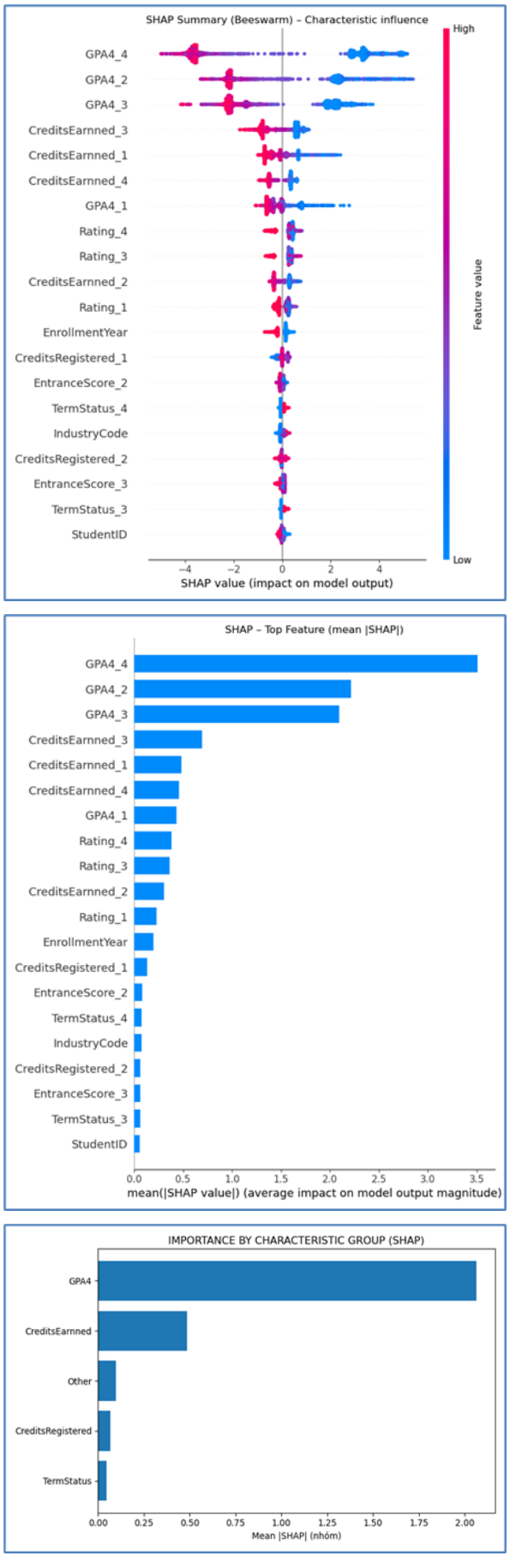
3.3. SHAP analysis and feature importance

Based on the SHAP[10] analysis, the most influential features contributing to student dropout risk are identified through three SHAP visualizations.

Among these features, the semester grade point average (GPA4) plays the most critical role. High SHAP values associated with these features indicate their strong impact on the model's decision-making process regarding dropout risk. Specifically, GPA4 exhibits the largest mean

impact, with a SHAP value of 2.062670 (Figure 7).

Figure 7. SHAP Analysis Results with Features



The number of earned credits (CreditsEarned) also exhibits a notable influence, with a SHAP value of 0.485200. This result indicates that students' academic commitment, as reflected by credit completion, plays a significant role in predicting dropout likelihood.

In addition, the number of registered credits (CreditsRegistered) and the entrance score (EntranceScore) serve as auxiliary factors with a comparatively smaller impact than GPA4 and earned credits. Their corresponding SHAP values are 0.068841 and 0.053038, respectively, indicating that while these features contribute to the model's predictions, their overall influence is relatively limited.

Finally, academic warnings (Rating) exhibit a smaller yet non-negligible contribution, with a SHAP value of 0.045942, highlighting that academic warning indicators serve as direct signals of dropout risk.

4. APPLICATIONS

The research outcomes were piloted at the Faculty of Information Technology, Quy Nhon University, with the objective of early identification of student dropout risk using the XGBoost machine learning model. The model incorporates features such as semester grade point averages along with academic warning indicators to estimate the probability of student dropout. Specifically, students with a predicted dropout probability greater than 0.7 are issued early warnings, enabling academic advisors to intervene in a timely manner. Accordingly, targeted support measures can be implemented, including personalized academic advising and career guidance, enhancement of learning skills, as well as financial and psychosocial support. The application of the XGBoost model not only provides technical value but also carries significant implications for data-driven management, enabling the optimization of student monitoring and support processes. This model contributes to the development of an intelligent education ecosystem at Quy Nhon University, promoting sustainable and efficient practices in educational management.

From an implementation perspective, the proposed XGBoost-based system can be readily integrated into existing academic management information systems or learning management systems as a decision-support module. Owing to its moderate computational complexity and reliance on routinely collected academic data, the model is suitable for real-time or periodic

deployment at scale across higher education institutions.

5. DISCUSSION

The study results confirm that XGBoost is an effective model for the early prediction of students at risk of dropout, offering high accuracy and clear interpretability. XGBoost outperforms Logistic Regression due to its superior handling of nonlinear relationships, enabled by its boosting mechanism and strong regularization. Compared with Random Forest, XGBoost also achieves higher performance by sequentially combining multiple trees, thereby improving predictive accuracy.

However, this study has several notable limitations. The data used are restricted to two enrollment cohorts (2020–2021) at Quy Nhon University, which may limit the model's generalizability. In addition, the study does not consider students' behavioral and psychosocial factors, such as learning motivation, engagement, and stress, which may significantly influence dropout likelihood. Furthermore, data from online learning platforms (LMS, e-learning) have not been integrated, resulting in the model lacking certain important information.

Future research will focus on expanding the dataset to validate generalizability, integrating learning behavior data from LMS and e-learning platforms, and incorporating psychosocial factors to enhance the model's comprehensiveness.

6. CONCLUSION

This study successfully developed an XGBoost model to predict early student dropout at Quy Nhon University, achieving superior performance compared with Logistic Regression and Random Forest models. Early-stage academic features, such as end-of-term grade point averages, accumulated credits, and academic warning status, have been identified as key factors directly influencing students' risk of dropout. The study results provide robust empirical evidence on the effectiveness of machine learning techniques in higher education management, enabling education administrators to identify and intervene promptly with students at high risk of dropout. The application of this predictive model not only enhances educational quality but also contributes to the development of an early warning system, enabling educational institutions to implement more effective learning support and advisory interventions. This model opens opportunities for the development of decision support tools in higher education, aiming to optimize student support and enhance the quality of academic programs.

Acknowledgment

This research is conducted within the framework of science and technology projects at institutional level of Quy Nhon University under the project code T2025.848.19

REFERENCES

1. L. T. M. Le, T. T. A. Tran. The overview of the reality and the factors affecting university dropouts, *VNUHCM Journal of Social Sciences and Humanities*, **2024**, *8(3)*, 2683-2706.

2. D. W. Hosmer, S. Lemeshow, R. X. Sturdivant. *Applied Logistic Regression (3rd edition)*, Wiley, New Jersey (US), **2013**.

3. L. Breiman. Random Forests, *Machine Learning*, **2001**, *45(1)*, 5–32.

4. S. Ray. *A quick review of machine learning algorithms*, International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, **2019**.

5. T. Chen, C. Guestrin. *XGBoost: A Scalable Tree Boosting System*, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, USA, **2016**.

6. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin. *CatBoost: Unbiased boosting with categorical features*, International Conference on Neural Information Processing Systems, 32nd Montréal Canada, **2018**.

7. A. Tharwat. Classification assessment methods, *Applied Computing and Informatics*, **2021**, *17(1)*, 168–192.

8. D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve, *Machine Learning*, **2009**, *77(1)*, 103–123.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research*, **2002**, Vol.*16*, 321–357.
10. S. M. Lundberg, S.-I. Lee. *A unified approach to interpreting model predictions*, Conference on Neural Information Processing Systems (NeurIPS), *31st,* Long Beach, USA, **2017**.