BAREKENG: Journal of Mathematics and Its Applications
December 2025 Volume 19 Issue 4 Page 2679–2692
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss4pp2679-2692

THE EFFECT OF SAMPLE SIZE ON THE STABILITY OF XGBOOST MODEL PERFORMANCE IN PREDICTING STUDENT STUDY PERIOD

Muhammad Lintang Damar Sakti[1*], Jailani[2], Heri Retnawati[3], Kana Hidayati[4], Nur Hadi Waryanto[5], Zulfa Safina Ibrahim[6], Asma' Khoirunnisa[7], Firdaus Amruzain Satiranandi Wibowo[8], Miftah Okta Berlian[9], Angella Ananta Batubara[10]

[1,2,3,4,5,6,7,8,9,10]*Department of Mathematics Education, Faculty of Science and Mathematics, Universitas Negeri Yogyakarta*
Jln. Colombo No. 1, Yogyakarta, 55281, Indonesia

*Corresponding author's e-mail: *muhammadlintang.2021@student.uny.ac.id*

Article History:

Received: 14th February 2025
Revised: 3rd June 2025
Accepted: 16th June 2025
Available online: 1st September 2025

Keywords:

Bootstrap;
Sample Size;
Stability;
Study Period;
XGBoost.

ABSTRACT

Student success can be defined based on the period of study taken until graduation from college. Machine learning can be used to predict the factors that are thought to influence student success. To achieve optimal machine learning model performance, attention is needed on the sample size. This study aims to determine the effect of student sample size on the stability of model performance to predict student success. This research is quantitative. The data used is student data from a university in Yogyakarta from 2014 to 2019, totaling 19061 students. The target variable is the student study period in months, while the predictor variables are college entrance pathways, GPA from semester 1 to semester 6, and family socioeconomic conditions based on the father's and mother's income. This research uses the XGBoost model with the best hyperparameters and the bootstrap approach. Bootstrapping was performed on the original data by sampling twenty different sample sizes: 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000, 4250, 4500, 4750, and 5000. The resulting bootstrap samples were replicated ten times. Model performance evaluation uses the Root Mean Square Error (RMSE) value. The result of this research is the XGBoost model with the best hyperparameters, obtained through the training data division scheme of 90% and testing data of 10%, which has the smallest RMSE value of 8.318. The model uses the best hyperparameters: n_estimators of 75, max_depth of 8, min_child_weight of 5, eta of 0.07, gamma of 0.2, subsample of 0.8, and colsample_bylevel of 1. The XGBoost model with optimal hyperparameters demonstrates peak performance stability at a sample size of 1750 students, as evidenced by consistent RMSE values across 10 bootstrap replications, confirming that this data quantity provides the ideal balance between prediction accuracy and stability for estimating study duration.

How to cite this article:

M. L. D. Sakti, Jailani, H. Retnawati, K. Hidayati, N. H. Waryanto, Z. S. Ibrahim, A. Khoirunnisa, F. A. S. Wibowo, M. O. Berlian, and A. A. Batubara, "THE EFFECT OF SAMPLE SIZE ON THE STABILITY OF XGBOOST MODEL PERFORMANCE IN PREDICTING STUDENT STUDY PERIOD," *BAREKENG: J. Math. & App.*, vol. 19, iss. 4, pp. 2679-2692, December, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

The success of student studies is measured based on the study period in months taken by students to graduate from college. This measure of study success provides an overview of the duration required by each student to complete their study program. The measure is discrete, represented as a count of months, which is an integer. Integers are discrete variables [1].

The success of student studies, measured by the study period, can be influenced by several factors. Machine learning can be used to predict the study period of students. In addition, machine learning has the capability to build models from large datasets [2]. A model that can be used to predict the study period of students is XGBoost.

Xtreme Gradient Boosting (XGBoost) is an enhancement of gradient tree boosting that can effectively be used in large-scale machine learning cases [3]. XGBoost can be utilized to prevent overfitting [3]. Research conducted by [4] also stated that XGBoost provides better performance than RGBM and is faster than scikit-learn.

XGBoost was selected for this study due to several key advantages. First, as an enhancement of the Gradient Boosted Decision Trees (GBDT) algorithm, XGBoost overcomes the limitations of the base model through the addition of L1/L2 regularization that reduces overfitting [3]. Second, its ability to automatically handle missing data is particularly relevant given the frequently incomplete nature of educational data. Third, its parallel computing optimization enables efficient processing of large datasets [4], whereas similar algorithms like Random Forest prove less optimal for datasets of this scale.

This study divided the data into four schemes: the first scheme used 60% training data and 40% testing data, the second scheme used 70% training data and 30% testing data, the third scheme used 80% training data and 20% testing data, and the fourth scheme used 90% training data and 10% testing data.

To achieve optimal model performance, it can be evaluated using the Root Mean Square Error (RMSE). XGBoost requires proper hyperparameter tuning to improve prediction accuracy and assist in decision-making regarding factors influencing student study duration. Meanwhile, achieving optimal model performance in predicting student study duration also necessitates attention to sample size, which is suspected to affect model performance.

Larger sample sizes can help improve model performance by reducing variance and increasing prediction accuracy. Several studies have examined how sample size affects the stability of model performance. This research utilized one of the machine learning models, mixed-effects regression, and Monte Carlo simulation. The study concluded that a large sample size is essential for achieving stable model performance results.

Simulation of a model to replicate data as desired can be performed using bootstrapping. The bootstrap algorithm works by taking multiple independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of 0 using the empirical standard deviation of those replications [5]. Bootstrapping is conducted to learn about the model from a data distribution [6].

Research on bootstrap has been conducted by [7] regarding logistic regression models for classifying the graduation time of STIKOM Bali students using the bootstrap aggregating or bagging method. The researchers applied bagging to improve classification accuracy and parameter stability in the logistic regression model. The study results showed that the bagging approach in logistic regression increased classification accuracy by 1.01%. The best classification, at 86.40%, was achieved with 70 bootstrap replications.

Therefore, the researcher aims to examine the stability of model performance, observed through the Root Mean Square Error (RMSE) value, in predicting student study duration using the XGBoost model with the bootstrap method. If it is proven that the number of student sample sizes affects the stability of model performance in predicting study duration, it can be concluded that sample size influences model performance stability. Research to determine the stability of model performance for predicting student study duration needs to consider student sample sizes.

2. RESEARCH METHODS

2.1 Data Collection

This research uses secondary data. The data in this study are student data from 2014 to 2019, totalling 19061 students. The population in this study was students at one of the universities in Yogyakarta.

2.2 Variable Definition

The variables used in this study are the period of study of students in units of months taken by students from the beginning of entering college to graduation, college entrance paths consist of SNBP (National Selection Based on Achievement), SNBT (National Selection Based on Test), and Independent Selection, father's income, mother's income, GPA (Grade Point Average) for semester 1, semester 2, semester 3, semester 4, semester 5, and semester 6.

Table 1. Definition of Variable

Variable	Type	Description of Variables
Study Period	Continuous	42- 84 months
GPA for semester 1	Continuous	0.00 – 4.00
GPA for semester 2	Continuous	0.00 – 4.00
GPA for semester 3	Continuous	0.00 – 4.00
GPA for semester 4	Continuous	0.00 – 4.00
GPA for semester 5	Continuous	0.00 – 4.00
GPA for semester 6	Continuous	0.00 – 4.00
College Entrance Pathways	Categorical	1 : SNBT 2 : SNBP 3 : Independent Selection
Father's Income	Categorical	1 : Rp0 2 : Rp1,000 - Rp500,000 3 : Rp500,000 - Rp1,000,000 4 : Rp1,001,000 - Rp1,500,000 5 : Rp1,501,000 - Rp2,000,000 6 : Rp2,001,000 - Rp2,500,000 7 : Rp2,501,000 - Rp3,000,000 8 : Rp3,001,000 - Rp3,500,000 9 : Rp3,501,000 - Rp4,000,000 10 : > Rp4,000,000
Mother's Income	Categorical	1 : Rp0 2 : Rp1,000 - Rp500,000 3 : Rp501,000 - Rp1,000,000 4 : Rp1,001,000 - Rp1,500,000 5 : Rp1,501,000 - Rp2,000,000 6 : Rp2,001,000 - Rp2,500,000 7 : Rp2,501,000 - Rp3,000,000 8 : Rp3,001,000 - Rp3,500,000 9 : Rp3,501,000 - Rp4,000,000 10 : > Rp4,000,000

1) Study Period

According to [8], the success of student studies is defined by students completing their education on time. The standard duration for undergraduate students to complete their studies on time is four years, equivalent to eight semesters.

Several factors influence the length of a student's study period. Research conducted by [9] on factors presumed to affect academic achievement and the length of study found two factors influencing student performance: gender and university admission pathways. Meanwhile, factors influencing the length of study are gender, type of higher education institution, Grade Point Average (GPA), and university admission pathways.

Research conducted by [8] on the classification of factors affecting the length of student study identified presumed factors such as university admission pathways, gender, first-semester GPA, region of

origin, and family economic conditions. The study concluded that the factors influencing the length of student study are first-semester GPA and family economic conditions.

2) Semester GPA

Regulation of the Minister of Education, Culture, Research and Technology Number 53 of 2023 concerning Quality Assurance of Higher Education states that the results of learning outcomes each semester are expressed in the Semester GPA. Assessment of student learning outcomes in courses is expressed in an achievement index or a description of passing and not passing (Regulation of the Minister of Education, Culture, Research and Technology Number 53 of 2023 concerning Quality Assurance of Higher Education). The form of achievement index assessment is expressed in letter A equivalent to number 4, letter B equivalent to number 3, letter C equivalent to number 2, letter D equivalent to number 1, and letter E equivalent to number 0 (Regulation of the Minister of Education, Culture, Research and Technology Number 53 of 2023 concerning Quality Assurance of Higher Education).

3) College Entrance Pathways

Ministerial Regulation of Education, Culture, Research, and Technology Number 48 of 2022 on the Admission of New Students for Diploma and Undergraduate Programs at State Universities stipulates that the scope of new student admissions for diploma and undergraduate programs at State Universities (PTN) includes diploma three, diploma four or applied undergraduate, and undergraduate programs. New student admissions are conducted through national selection based on achievements (SNBP), national selection based on tests (SNBT), and independent selection by universities (Ministerial Regulation of Education, Culture, Research, and Technology Number 48 of 2022 on the Admission of New Students for Diploma and Undergraduate Programs at State Universities).

4) Socio-economic Condition

According to **[10]**, socioeconomic conditions can be observed through the income received by each individual in a family from their employment. Income exceeding consumption levels indicates family welfare, whereas income below consumption levels reflects a lack of welfare for the family **[10]**. Community welfare can be assessed from economic and social perspectives, such as income levels, monthly expenditures for food and non-food, production levels, investments, and others. From a social perspective, it includes education levels, work ethic, type of occupation, population dynamics, and more.

Learning motivation can be influenced by the family environment, particularly the socioeconomic status of parents, which supports students' academic achievement. Student learning outcomes, which require support to facilitate their education, are influenced by their parents' social conditions **[11]**. Separating fathers' and mothers' incomes allows for the identification of each parent's specific contribution to the family's socioeconomic condition.

2.3 Analytical Approach

The model used in this study is XGBoost. In this study, the Bootstrapping method will be used in the data resampling process. The analysis steps using R Studio are as follows.

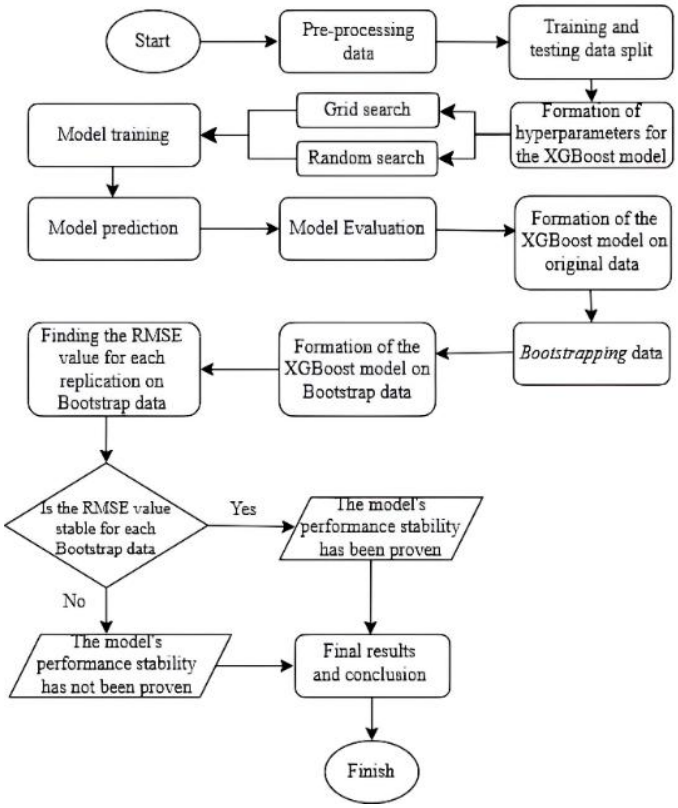

Figure 1. Research Flow Diagram

Based on **Figure 1**, this research began with the preparation of student data to be used in the modeling process. The next step involved data pre-processing, which included cleaning the data, adjusting formats, and removing irrelevant records. Afterward, the dataset was split into training and testing sets. Hyperparameter tuning for the XGBoost model was then carried out using two approaches: grid search and random search, to identify the best combination of parameters. The XGBoost model was trained using the training data and then used to make predictions on the testing data. The model's performance was evaluated using the Root Mean Square Error (RMSE) metric, and the best-performing model was selected based on the lowest RMSE value. The best model was then applied to the original full dataset.

To assess the model's stability, a bootstrapping process was conducted using R to generate multiple replicated sample datasets. The XGBoost model was rebuilt on each of these bootstrap samples. Each model's performance was measured using RMSE for every replication. The RMSE values obtained were then analyzed to determine whether they were stable across replications. If the RMSE values did not differ significantly from one replication to another, it was concluded that the model's performance was stable for that particular sample size.

1) Machine Learning

Machine learning is an application of artificial intelligence that enables systems to learn automatically from a set of data to perform specific tasks without being explicitly programmed **[12]**. Machine learning can be defined as a computer application and mathematical algorithm adopted through learning from data to generate future predictions. The learning process in machine learning consists of two stages: training and testing. The training data is used to train the algorithm, while the testing data is used to evaluate the algorithm's performance on new, unseen data.

In machine learning, the quality of the resulting model is highly influenced by the quality of the training data used. According to **[13]**, most machine learning algorithms have settings called hyperparameters, which control the behavior of the algorithm. These hyperparameters are not adapted by the algorithm itself during the learning process. When the dataset is too small, an alternative procedure that allows the use of all examples in estimating the average test error is cross-validation **[13]**. Cross-validation splits the dataset into different subsets of training and testing data, randomly selected from the original dataset **[13]**.

2) Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a model first proposed by Tianqi Chen and Carlos Guestrin in 2011. According to [3], Extreme Gradient Boosting is an algorithm that can be used to solve various problems, particularly in regression, classification, and ranking. The XGBoost algorithm utilizes the concept of ensemble learning, which combines the results of multiple models to generate more accurate predictions [14]. XGBoost aggregates predictions from multiple weak learners, known as decision trees. It employs a regularized model to construct regression tree structures, reducing model complexity to prevent overfitting [3].

XGBoost addresses the limitations of traditional gradient boosting algorithms by introducing various innovations, such as regularization, sequential decision trees, and more effective handling of missing data. The core concept of this algorithm is to iteratively adjust learning parameters to minimize the loss function [3]. XGBoost tends to assign higher weights to rare classes in class imbalance scenarios because it focuses on improving model performance for the minority class, which has fewer samples [14]. In cases of class imbalance, errors in the minority class produce larger gradients, prompting the algorithm to assign higher weights to that class to minimize errors [14]. The XGBoost algorithm was briefly introduced by [4] as follows:

A tree model integrated with the addition method [4].

$$\widehat{y_t} = \sum_i f_i(x_i), f_i \epsilon F \tag{1}$$

Information:

$\widehat{y_t}^t$: Predicted value for the i-th data point after t iterations

$f_i(x_i)$: Base tree model

The objective function is as follows [4].

$$L = \sum_i l(\widehat{y_i}, y_i) + \sum_i \Omega(f_i) \tag{2}$$

Information:

l : Predicted value for the i-th data point after t iterations, loss function that represents the error between the prediction and the true value

The regularization function is as follows [4].

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda + ||w||^2 \tag{3}$$

Information:

γ : A parameter that controls the number of leaf nodes

T : Number of leaf nodes in the tree

λ : Parameter that controls the weight of leaf nodes

w : Weights of leaf nodes in the tree

The information gain of the objective function is as follows [4].

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\epsilon I_L} g_i)^2}{\sum_{i\epsilon I_L} h_L + \lambda} + \frac{(\sum_{i\epsilon I_R} g_i)^2}{\sum_{i\epsilon I_L} h_R + \lambda} + \frac{(\sum_{i\epsilon I_I} g_i)^2}{\sum_{i\epsilon I_L} h_I + \lambda}\right] \tag{4}$$

Information:

g_i : The gradient used to show how big the error is

h_i : Hessian, which is used to show how changes in the predicted value affect the error generated by the model.

3) Hyperparameter Tuning

The parameters that can improve model performance are called hyperparameters [3]. The methods used for hyperparameter tuning include Grid Search Cross-Validation (CV) and Random Search Cross-Validation (CV). Grid Search Cross Validation (Grid Search CV) is a hyperparameter optimization method in machine learning that works by constructing and evaluating every combination of predefined algorithm parameters within a grid. It conducts an exhaustive search to test all possible parameter combinations [15]. Random Search Cross Validation (Random Search CV) is a model selection method that assigns hyperparameter values by randomly selecting combinations of hyperparameters to train the model.

Below are the best hyperparameter values for the XGBoost model according to [16].

Table 2. Best Values for Hyperparameter Tuning

Hyperparameter	The Usefulness of Hyperparameters
n_estimators	The number of trees used for prediction
max_depth	Setting the maximum depth of each tree
min_child_weight	The minimum weight required for each tree branch
eta (learning_rate)	Helps regulate the magnitude of changes applied
gamma	Sets the minimum reduction in loss
subsample	Controls how much data is used in each iteration
colsample_bylevel	Determines the percentage of training data used to build the tree

Table 2 presents the best values selected for hyperparameter tuning in the XGBoost model, along with an explanation of their usefulness. The n_estimators parameter represents the number of trees used in the boosting process, which influences both the model's performance and computational cost. The max_depth controls the maximum depth of each tree, affecting the model's ability to capture complex patterns in the data. The min_child_weight sets the minimum weight required to create a new child node, helping to prevent overfitting by requiring sufficient data in each leaf. The eta (learning_rate) determines the step size in the learning process, where smaller values typically lead to better generalization but require more trees. The gamma parameter specifies the minimum reduction in loss needed to make a further split, acting as a regularization mechanism to control tree complexity. The subsample parameter defines the proportion of training data used in each iteration, introducing randomness to avoid overfitting. Lastly, colsample_bytree controls the fraction of features used to build each tree, which also helps improve the model's generalization capability. These hyperparameters play a crucial role in optimizing model performance and ensuring robust predictions.

4) Model Evaluation

Model evaluation is used to select the best model and can be performed using the Root Mean Squared Error (RMSE). According to [17], the best model is the one with the lowest RMSE among the built models. The RMSE equation is as follows [18].

$$RMSE = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)} \tag{5}$$

Information:

y_i : Actual value of the i-th observation $(i = 1,2,\ldots,n)$

$\hat{y}_i$: Predicted value of the i-th observation $(i = 1,2,\ldots,n)$

n : Number of observation samples

5) Bootstrapping

Bootstrap is a method that involves sampling with replacement (resampling) from observed data [6]. According to [5], bootstrap is a widely applied resampling method that enables the creation of more realistic models. The purpose of using the bootstrap method is to address issues related to small datasets, data that deviate from assumptions, or data that do not follow any specific distribution assumptions [19]. Research on

model performance stability requires repeated resampling from the original data to assess the errors produced by each sample. Therefore, the bootstrap method is essential in such studies.

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Data Description

The target variable used in this study is the students' study duration, measured in months, from their initial enrollment in university until graduation between 2014 and 2019. The predictor variables suspected to influence study duration include university admission pathways, family economic conditions (father's and mother's income), and students' academic performance, represented by their Grade Point Average (GPA) from the first to the sixth semester.

According to [9], factors influencing students' study duration include GPA and university admission pathways. Meanwhile, [8] found that study duration is affected by GPA and family economic conditions, specifically parents' income. The characteristics of these variables are presented in Table 3 and Table 4.

Table 3. Descriptive Statistics for Continuous Variables

Variable	Mean	Variance	Minimum	Maximum
Study Period	49.03	81,46	42	84
GPA for semester 1	3.11	1.12	0.00	4.00
GPA for semester 2	3.44	0.06	0.00	4.00
GPA for semester 3	3.46	0.08	0.00	3.98
GPA for semester 4	3.47	0.08	0.00	3.98
GPA for semester 5	3.50	0.07	0.00	3.98
GPA for semester 6	3.51	0.08	0.00	3.97

Based on Table 3, the average study duration for students in higher education is 49.03 months, or approximately 8 semesters, indicating that most students complete their studies on time. Additionally, the study duration has a variance of 81.45, suggesting a considerable degree of variability. A higher variance implies greater diversity in students' study durations. According to [20], the larger the difference between the highest and lowest values in the data, the greater the variation.

Table 3 also shows variability in students' Grade Point Averages (GPA), but overall, students demonstrate good academic performance, as the average GPA in each semester exceeds 3.00. Furthermore, the average GPA consistently increases across semesters. A study conducted by [21] examined factors suspected to influence GPA. The results indicated that the factors affecting students' GPA include study hours at home, study hours on campus, and the number of organizations they participate in.

Table 4. Descriptive Statistics for Categorical Variables

Variable	Category	Count
College Entrance Pathways	1 : SNBT	6003
	2 : SNBP	7235
	3 : Independent Selection	5823
Father's Income	1 : Rp0	2371
	2 : Rp1,000 - Rp500,000	1895
	3 : Rp501,000 - Rp1,000,000	3571
	4 : Rp1,001,000 - Rp1,500,000	2827
	5 : Rp1,501,000 - Rp2,000,000	1573
	6 : Rp2,001,000 - Rp2,500,000	1017
	7 : Rp2,501,000 - Rp3,000,000	1203
	8 : Rp3,001,000 - Rp3,500,000	959
	9 : Rp3,501,000 - Rp4,000,000	1373
	10 : > Rp4,000,000	2272
Mother's Income	1 : Rp0	9631
	2 : Rp1,000 - Rp500,000	2726
	3 : Rp501,000 - Rp1,000,000	1934
	4 : Rp1,001,000 – Rp1,500,000	940

Variable	Category	Count
	5 : Rp1,501,000 - Rp2,000,000	485
	6 : Rp2,001,000 - Rp2,500,000	446
	7 : Rp2,501,000 - Rp3,000,000	606
	8 : Rp3,001,000 - Rp3,500,000	577
	9 : Rp3,501,000 - Rp4,000,000	894
	10 : > Rp4,000,000	922

Based on **Table 4**, there is diversity in students' admission pathways into higher education, including SNBT, SNBP, and Independent Selection. Additionally, **Table 4** reflects the economic conditions of students' families, as seen from the father's and mother's income, providing insights into the different economic backgrounds of students. According to **Table 4**, the Seleksi Nasional Berdasarkan Prestasi (SNBP) is the most common university admission pathway, with 7235 students. Meanwhile, the Independent Selection pathway has the fewest students, with 5823 students.

Regarding fathers' income, out of ten income categories, Category 3 (Rp501,000 - Rp1,000,000) has the highest number of fathers, totaling 3571. In contrast, Category 8 (Rp3,001,000 - Rp3,500,000) has the fewest, with only 959 fathers. For mothers' income, the most common category is Category 1 (Rp0 or no income), with 9631 mothers. On the other hand, Category 6 (Rp2,001,000 - Rp2,500,000) has the fewest mothers, totaling 446.

3.1.2 XGBoost Model with Hyperparameter Tuning

This study uses two testing schemes. The first testing scheme is the XGBoost model testing with hyperparameter tuning using grid search cross-validation (CV), and the second testing scheme uses random search cross-validation (CV). According to several previous studies, the best model is obtained with the training and testing data splitting scheme, as shown in **Table 5**.

Table 5. Data Splitting Scheme

Researcher	Data Splitting
Wicaksono et al. (2024) [22]	Training: 60% Testing : 40%
Rayadin et al. (2024) [23]	Training: 70% Testing : 30%
Wibowo (2022) [24]	Training: 80% Testing : 20%
Saputra et al. (2024) [25]	Training: 90% Testing : 10%

To obtain the best model, it is necessary to use the optimal data splitting scheme for training and testing, as shown in **Table 5**. Additionally, achieving the best model also requires using the optimal hyperparameters. According to previous research, the best XGBoost model can be obtained with the hyperparameter values listed in **Table 6**.

Table 6. Hyperparameter Value Scheme

Researcher	Hyperparameter	Hyperparameter Values
Syukron et al. (2020) [26]	n_estimators	(30; 50; 75; 100; 125)
Jange (2022) [27]	max_depth	(8; 10; 12; 15)
Syukron et al. (2020) [26]	min_child_weight	(1; 2; 3; 4; 5)
Agustin et al. (2023) [28]	eta (learning_rate)	(0.07; 0.16; 0.21; 0.43)
Wibowo (2022) [24]	gamma	(0.001; 0.01; 0.1; 0.2; 0.3)
Rayadin et al. (2024) [23]	subsample	(0.6; 0.7; 0.8)
Febriantoro et al. (2023) [3]	colsample_bylevel	(0.1; 0.2; 0.25; 1)

To optimize the predictive model in this study, K-Fold Cross-Validation can be utilized. According to [29], K-Fold Cross-Validation can help reduce computational time while maintaining the accuracy of the prediction process. Additionally, K-Fold Cross-Validation can address issues of overfitting and underfitting. Based on previous research, the best XGBoost model can be obtained by determining the appropriate K-Fold, as shown in **Table 7**.

Table 7. K-Fold Cross-Validation Scheme

Researcher	K-Fold
Syukron et al. (2020) [26]	5-fold CV
Pardede & Nurrohmah (2024) [30]	5-fold CV

3.1.3 Best Model Selection

The best model selection is based on the smallest RMSE value from each model built using various testing schemes. The best model in this study is the XGBoost model, using a data splitting scheme of 90% for training and 10% for testing, with grid search cross-validation (CV), achieving the lowest RMSE value of 8.318. The optimal hyperparameter values for this model can be seen in **Table 8**.

Table 8. Best Model Hyperparameters

Hyperparameter	Best Hyperparameters
n_estimators	75
max_depth	8
min_child_weight	5
eta (learning_rate)	0.07
gamma	0.2
subsample	0.8
colsample_bylevel	1

3.1.4 Bootstrapping

Bootstrapping is performed by sampling with replacement for a total of 10 replications, using sample sizes of 250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000, 4250, 4500, 4750, and 5000. For each sample, the best XGBoost model is built based on the optimal data splitting scheme for training and testing, as well as the previously determined hyperparameters. The RMSE values for each sample can be seen in **Figure 2**.

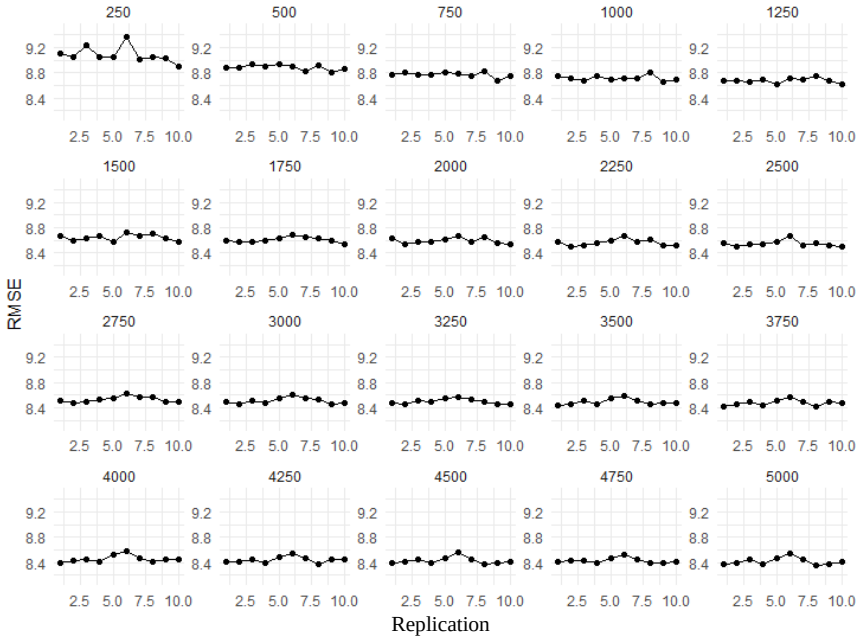

Figure 2. Model Performance Stability

Based on **Figure 2**, the most stable sample size, as indicated by the RMSE values obtained across 10 replications, is 1750 data samples. Research on model performance stability has been conducted, revealing that the number of student samples influences both the RMSE values and the stability of the XGBoost model's performance using the optimal hyperparameters.

3.2 Discussion

This study utilizes data from students of a university in Yogyakarta from 2014 to 2019, totaling 19061 records. The predictor variable used is the students' study duration in months, while the response variables include university admission pathways, GPA from the first to the sixth semester, and the family's socioeconomic status based on the father's and mother's income. The objective of this study is to assess the stability of the model's performance in predicting students' study duration by considering different student sample sizes using the XGBoost model.

Based on the XGBoost model with the optimal hyperparameters, the best model for predicting students' study duration uses a data split of 90% for training and 10% for testing, achieving the lowest RMSE value of 8.318. The model's best hyperparameters, as listed in **Table 8**, are as follows: n_estimators = 75, max_depth = 8**,** min_child_weight = 5, eta **(**learning rate**)** = 0.07, gamma = 0.2, subsample = 0.8, and colsample_bylevel = 1.

Based on the RMSE plot of the bootstrapped sample data in **Figure 2**, the sample size of 250 appears unstable due to the significant differences in RMSE values across replications. For the 250-sample data, the highest RMSE value is 9.243, while the lowest is 8.908. Similarly, sample sizes of 500, 750, 1000, 1250, 1500, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000, 4250, 4500, 4750, and 5000 also exhibit instability. However, based on the RMSE plot in Figure 2, the sample size of 1750 with ten replications is the most stable, as it has the smallest RMSE variation across replications.

Thus, this study concludes that the stability of the XGBoost model with the optimal hyperparameters for predicting study duration is achieved with a sample size of 1750 students**.** Research on model performance stability has been conducted, showing that model performance depends on sample size. These findings are supported by **[31]**, who used different methods and datasets, and also found that sample size affects model performance stability.

This study is limited to data from a single university in Yogyakarta, which may affect the generalizability of its findings. The authors recommend future research to incorporate data from multiple institutions across different regions and with varied characteristics to enhance result validity. Such expansion would yield more robust and representative insights into student study duration patterns in Indonesia's diverse higher education landscape.

4. CONCLUSION

The model was optimized through grid search cross-validation (CV) and 5-fold CV, achieving the lowest RMSE value of 8.318. The optimal hyperparameters for this model include n_estimator**s** = 75, max_depth = 8, min_child_weight = 5, eta (learning rate) = 0.07, gamma = 0.2, subsample = 0.8, and colsample_bylevel = 1. The sample size in this study, which represents the number of students, influences the model's performance in predicting students' study duration. Performance is evaluated using the RMSE value, which indicates that as the sample size increases, the RMSE decreases. This suggests that the model performs better when trained with a larger dataset, as it can generalize more effectively and provide more accurate predictions. The stability of the XGBoost model's performance with the best hyperparameters is observed at a sample size of 1750 students. This conclusion is based on RMSE values obtained across different sample sizes after 10 replications using bootstrap data. The model remains consistent and reliable at this sample size, indicating that 1750 data provide an optimal balance between accuracy and stability in predicting students' study duration.

AUTHOR CONTRIBUTIONS

Muhammad Lintang Damar Sakti: Conceptualization, Data curation, Investigation, Methodology, Writing - original draft. Jailani: Conceptualization, Funding acquisition, Supervision, Validation, Writing - original draft. Heri Retnawati: Conceptualization, Funding acquisition, Resources, Supervision, Validation, Writing - original draft, Writing - review & editing. Kana Hidayati: Conceptualization, Funding acquisition, Supervision, Validation, Writing - original draft. Nur Hadi Waryanto: Conceptualization, Funding acquisition, Software, Supervision, Validation. Zulfa Safina Ibrahim: Investigation, Methodology, Writing -

original draft, Visualization. Asma' Khoirunnisa: Investigation, Methodology, Writing - original draft. Firdaus Amruzain Satiranandi Wibowo: Investigation, Methodology, Writing - original draft. Miftah Okta Berlian: Investigation, Methodology, Writing - original draft. Angella Ananta Batubara: Investigation, Methodology, Writing - original draft. All authors discussed the results and contributed to the final manuscript.

FUNDING STATEMENT

This research received no funding from any public agency.

ACKNOWLEDGMENT

The authors would like to thank all those who have participated and provided support in this research.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest regarding this study.

REFERENCES

[1] A. Sugianto, "TYPES OF DATA VARIABLES (DISCRETE AND CONTINUOUS VARIABLES)," 2016, [Online]. Available: https://adoc.pub/jenis-jenis-data-variabel-variabel-diskrit-dan-variabel-kont.html

[2] E. Febriantoro, E. Setyati, and J. Santoso, "MODELING TOY SALES QUANTITY PREDICTION USING LIGHT GRADIENT BOOSTING MACHINE," *SMARTICS J.*, vol. 9, no. 1, pp. 7–13, 2023, doi: https://doi.org/10.21067/smartics.v9i1.8279.

[3] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "APPLICATION OF EXTREME GRADIENT BOOSTING (XGBOOST) METHOD IN CREDIT CARD CUSTOMER CLASSIFICATION," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: https://doi.org/10.31605/jomta.v4i1.1792.

[4] T. Chen and C. Guestrin, "XGBOOST: A SCALABLE TREE BOOSTING SYSTEM," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016, doi: https://doi.org/10.1145/2939672.2939785.

[5] B. Efron and R. J. Tibshirani, *AN INTRODUCTION TO THE BOOTSTRAP*. London: Chapman and Hall, Inc, 1993. doi: https://doi.org/10.1007/978-1-4899-4541-9.

[6] F. P. A. putra Rachman, R. Goejantoro, and M. N. Hayati, "DETERMINATION OF THE NUMBER OF BOOTSTRAP REPLICATIONS USING THE PRETEST METHOD IN THE INDEPENDENT SAMPLE T-TEST (REGIONAL ORIGINAL REVENUE OF REGENCIES/CITIES IN EAST KALIMANTAN AND NORTH KALIMANTAN PROVINCES IN 2015)," *J. Eksponensial*, vol. 9, no. 1, pp. 35–40, 2018, [Online]. Available: blob:chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/27956194-298e-41aa-9c94-ec7a856961d4

[7] I. K. P. Suniantara, G. Suwardika, and I. G. A. Astapa, "BAGGING LOGISTIC REGRESSION FOR IMPROVING THE CLASSIFICATION ACCURACY OF STUDENT GRADUATION TIME AT STIKOM BALI," *J. Din.*, vol. 09, no. 1, pp. 10–19, 2018.

[8] Darwin and S. Zurimi, "APPLIED MODEL ANALYSIS OF MULTIVARIATE ADAPTIVE REGRESSION SPLINE (MARS) ON THE CLASSIFICATION OF FACTORS AFFECTING THE STUDY PERIOD OF FKIP UNIVERSITAS DARUSSALAM AMBON STUDENTS," *J. SIMETRIK*, vol. 9, no. 2, pp. 250–255, 2019.doi ; https://doi.org/10.31959/js.v9i2.426

[9] Nalim, H. L. Dewi, and M. A. Safii, "ANALYSIS OF FACTORS INFLUENCING STUDENTS' academic success at PTKIN in Jawa Tengah province," *J. Has. Penelit. dan Kaji. Kepustakaan di Bid. Pendidikan, Pengajaran dan Pembelajaran*, vol. 7, no. 4, pp. 1003–1013, 2021, doi: https://doi.org/10.33394/jk.v7i4.3430.

[10] N. Hanum and S. Safuridar, "ANALYSIS OF FAMILY SOCIO-ECONOMIC CONDITIONS ON FAMILY WELFARE IN GAMPONG KARANG ANYAR, LANGSA," *J. Samudra Ekon. dan Bisnis*, vol. 9, no. 1, pp. 42–49, 2018, doi: https://doi.org/10.33059/jseb.v9i1.460.

[11] B. Sudarwanto, "THE INFLUENCE OF PARENTS' SOCIO-ECONOMIC STATUS AND LEARNING MOTIVATION ON THE ACADEMIC ACHIEVEMENT OF STUDENTS AT SMPN 4 WONOSOBO," *Media Manaj. Pendidik.*, vol. 1, no. 1, p. 116, 2018, doi: 10.30738/mmp.v1i1.2881.

[12] B. Mahesh, "MACHINE LEARNING ALGORITHMS," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: https://doi.org/10.21275/ART20203995.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *DEEP LEARNING*. Cambridge: The MIT Press, 2016. [Online]. Available: http://deeplearning.net/

[14] E. J. Sudarman and S. Budi, "DEVELOPMENT OF AN EXTREME GRADIENT BOOSTING MACHINE INTELLIGENCE MODEL FOR PREDICTING STUDENT ACADEMIC SUCCESS," *J. Strateg.*, vol. 5, no. 2, pp. 297–314, 2023.

[15] L. N. Aina, V. R. S. Nastiti, and C. S. K. Aditya, "IMPLEMENTATION OF EXTRA TREES CLASSIFIER WITH GRID SEARCH CV OPTIMIZATION FOR ADAPTATION LEVEL PREDICTION," *MIND (Multimedia Artif. Intell. Netw. Database) J.*, vol. 9, no. 1, pp. 78–88, 2024, [Online]. Available: https://ejurnal.itenas.ac.id/index.php/mindjournal/article/view/10899

[16] B. Quinto, *NEXT-GENERATION MACHINE LEARNING WITH SPARK*. 2020. doi: ttps://doi.org/10.1007/978-1-4842-5669-5.

[17] N. N. Pandika Pinata, I. M. Sukarsa, and N. K. Dwi Rusjayanthi, "TRAFFIC ACCIDENT PREDICTION IN BALI USING XGBOOST IN PYTHON," *J. Ilm. Merpati (Menara Penelit. Akad. Teknol. Informasi)*, vol. 8, no. 3, p. 188, 2020, doi: https://doi.org/10.24843/JIM.2020.v08.i03.p04.

[18] T. O. Hodson, "ROOT-MEAN-SQUARE ERROR (RMSE) OR MEAN ABSOLUTE ERROR (MAE): WHEN TO USE THEM OR NOT," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022, doi: https://doi.org/10.5194/gmd-15-5481-2022.

[19] A. R. Naufal, R. Satria, and A. Syukur, "IMPLEMENTATION OF BOOTSTRAPPING FOR CLASS IMBALANCE AND WEIGHTED INFORMATION GAIN FOR FEATURE SELECTION IN THE SUPPORT VECTOR MACHINE ALGORITHM FOR CUSTOMER LOYALTY PREDICTION," *J. Intell. Syst.*, vol. 1, no. 2, pp. 98–108, 2015.

[20] M. Maswar, "DESCRIPTIVE STATISTICAL ANALYSIS OF ECONOMETRICS FINAL EXAM SCORES OF STUDENTS USING SPSS 23 & EVIEWS 8.1," *J. Pendidik. Islam Indones.*, vol. 1, no. 2, pp. 273–292, 2017, doi: https://doi.org/10.35316/jpii.v1i2.54.

[21] I. Veronika Girsang *et al.*, "ANALYSIS OF FACTORS INFLUENCING THE CUMULATIVE GRADE POINT AVERAGE (GPA) OF STUDENTS IN THE DEVELOPMENT ECONOMICS DEPARTMENT, FACULTY OF ECONOMICS AND BUSINESS, UNIVERSITAS PALANGKA RAYA," *J. Ilm. Mhs.*, vol. 2, no. 1, pp. 144–156, 2024, [Online]. Available: https://doi.org/10.59603/niantanasikka.v2i1.

[22] D. F. Wicaksono, R. S. Basuki, and D. Setiawan, "IMPLEMENTATION OF MACHINE LEARNING IN PREDICTING DISEASE CASE LEVELS IN INDONESIA," *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 736, 2024, doi: https://doi.org/10.30865/mib.v8i2.7501.

[23] M. A. Rayadin, M. Musaruddin, R. A. Saputra, and I. Isnawaty, "IMPLEMENTATION OF ENSEMBLE LEARNING USING XGBOOST AND RANDOM FOREST METHODS FOR PREDICTING BATTERY REPLACEMENT TIME," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, pp. 111–119, 2024.doi: https://doi.org/10.37148/bios.v5i2.128

[24] A. Wibowo, "EARTHQUAKE MAGNITUDE PREDICTION USING MACHINE LEARNING WITH THE XGBOOST MODEL AS A STRATEGIC STEP IN EARTHQUAKE-RESISTANT BUILDING STRUCTURE PLANNING IN INDONESIA," *MESA (Teknik Mesin, Tek. Elektro, Tek. Sipil …*, vol. 6, no. 1, pp. 18–29, 2022, [Online]. Available: http://www.ejournal.unsub.ac.id/index.php/FTK/article/view/1829

[25] A. S. Saputra, B. N. Sari, and C. Rozikin, "IMPLEMENTATION OF THE EXTREME GRADIENT BOOSTING (XGBOOST) ALGORITHM FOR CREDIT RISK ANALYSIS," *J. Ilm. Wahana Pendidik.*, vol. 10, no. 7, pp. 27–36, 2024, doi: 10.5281/zenodo.10960080.

[26] M. Syukron, R. Santoso, and T. Widiharih, "COMPARISON OF SMOTE RANDOM FOREST AND SMOTE XGBOOST METHODS FOR CLASSIFYING HEPATITIS C SEVERITY LEVELS ON IMBALANCED CLASS DATA," *J. GAUSSIAN*, vol. 9, no. 3, pp. 227–236, 2020.doi: https://doi.org/10.14710/j.gauss.v9i3.28915

[27] B. Jange, "PREDICTION OF BCA BANK STOCK PRICES USING XGBOOST," *Arbitr. J. Econ. Account.*, vol. 3, no. 2, pp. 231–237, 2022, doi: https://doi.org/10.47065/arbitrase.v3i2.495.

[28] E. Agustin, A. Eviyanti, and N. L. Azizah, "EPILEPSY DETECTION THROUGH EEG SIGNALS USING DWT AND EXTREME GRADIENT BOOSTING," *J. Media Inform. Budidarma*, vol. 7, no. 1, p. 117, 2023, doi: https://doi.org/10.30865/mib.v7i1.5412.

[29] Jimmy, L. D. Yulianto, E. H. Hermaliani, and L. Kurniawati, "PENERAPAN MACHINE LEARNING DALAM ANALISIS STADIUM PENYAKIT HATI UNTUK PROSES DIAGNOSIS DAN PERAWATAN," *RESOLUSI Rekayasa Tek. Inform. dan Inf.*, vol. 3, no. 4, pp. 170–180, 2023.

[30] J. Pardede and D. Nurrohmah, "HEPATITIS IDENTIFICATION USING BACKWARD ELIMINATION AND EXTREME GRADIENT BOOSTING METHODS," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 2, pp. 302–313, 2024, doi: https://doi.org/10.20473/jisebi.10.2.302-313.

[31] A. Pate, R. Emsley, M. Sperrin, G. P. Martin, and T. van Staa, "IMPACT OF SAMPLE SIZE ON THE STABILITY OF RISK SCORES FROM CLINICAL PREDICTION MODELS: A CASE STUDY IN CARDIOVASCULAR DISEASE," *Diagnostic Progn. Res.*, vol. 4, no. 1, 2020, doi: https://doi.org/10.1186/s41512-020-00082-3.