JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 3, June 2026, hlm 509 – 514 ISSN 2550-0201 (Online)
DOI: https://doi.org/ 10.33330/jurteksi.v12i3.4644
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

IMPLEMENTATION OF XGBOOST FOR PREDICTING STUDENT GRADUATION USING SIMULATED DATASET

Dewi Anggraeni[1*], Sri Rezeki Maulina Azmi[1]
[1]Information System, Universitas Royal
email: *dewianngraeni2024123@gmail.com

Abstract: Student graduation is an urgent matter that is an indicator of the success of a university in producing its learning output. Several factors influence student graduation such as GPA, attendance, late taking credits, and lack of student involvement in academic activities. The urgency of this research, universities need a method that is able to predict student graduation early so that it can provide academic intervention to students who have the potential to experience delays or fail to graduate. However, limited access to real academic data is often an obstacle in the development of predictive models, Therefore, this study aims to implement the XGBoost algorithm to predict student graduation based on several academic variables, namely the Cumulative Grade Point Average (GPA), the number of credits taken, the percentage of attendance, and the average grade of students. Model training using the XGBoost algorithm using a simulation dataset of 500 students who are labeled as graduating into two classes, namely passed and failed. The results of the study showed that the classification performance was very good with an accuracy value of 99.6%, Precision 99.7%, recall 99.4%.

Keywords: xgboost algorithm; data mining; student graduation

Abstrak: Kelulusan mahasiswa merupakan hal urgensi yang menjadi indikator keberhasilan sebuah perguruan tinggi dalam menghasilkan output pembelajarannya. Beberapa Faktor yang mempengaruhi kelulusan mahasiswa seperti IPK, kehadiran, keterlambatan pengambilan SKS, serta kurangnya keterlibatan mahasiswa dalam aktifitas akademik. Yang menjadi urgensi penelitian ini, Perguruan tinggi memerlukan suatu metode yang mampu memprediksi kelulusan mahasiswa secara dini sehingga dapat memberikan intervensi akademik kepada mahasiswa yang berpotensi mengalami keterlambatan atau tidak lulus. Namun, keterbatasan akses terhadap data akademik riil sering menjadi kendala dalam pengembangan model prediksi, Oleh karena itu, penelitian ini bertujuan mengimplementasikan algoritma XGBoost untuk memprediksi kelulusan mahasiswa berdasarkan beberapa variabel akademik, yaitu Indeks Prestasi Kumulatif (IPK), jumlah SKS yang ditempuh, persentase kehadiran, dan nilai rata-rata mahasiswa. Pelatihan model menggunakan algoritma XGBoost dengan menggunakan dataset simulasi 500 mahasiswa yang diberi label kelulusan menjadi dua kelas yaitu lulus dan tidak lulus. Hasil penelitian menunjukan bahwa performance klasifikasi yang sangat baik dengan nilai accurasi sebesar 99,6%, Precision 99,7%, recall 99,4%.

Kata kunci: algoritma xgbosst; kelulusan mahasiswa; penambangan data

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 3, June 2026, hlm 509 – 514 ISSN 2550-0201 (Online)
DOI: https://doi.org/ 10.33330/jurteksi.v12i3.4644
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUTION

Graduating on time is one of the main indicators of successful higher education. The graduation rate of students not only affects individual academic achievement but also impacts the accreditation of study programs, the reputation of universities, the effectiveness of using educational resources, and the key performance indicators (KPIs) set by the government. The quality of students in completing an 8-semester course can reflect the quality of the institution's accreditation.

However, there are still many universities facing the problem of a high number of students who can't finish their studies within the ideal study period. Graduation delays are generally influenced by various factors, such as low GPA, a limited number of credits completed, low attendance, declining academic performance, and a lack of student involvement in academic and organizational activities. As a result, students risk having to extend their study period, which leads to higher education costs, delays in entering the workforce, and lower efficiency in higher education administration.

In previous research, the application of several data classification methods to produce predictions, such as [1] stating that applying the C4.5 Decision Tree algorithm with the C4.5 decision tree method successfully predicted student graduation well. The main factor influencing it is GPA, followed by study program and gender. The model has a high accuracy of 85.34%. Then in the research [2] The Naïve Bayes method is effective in predicting students' graduation rates. This method not only makes it easier to determine students' graduation outcomes, but it can also predict the number of students who will graduate. The prediction accuracy from student graduation rate data reaches 98.33%. Meanwhile, in this study [3] The Random Forest algorithm is used to predict students graduating on time. The research results show that Random Forest has an accuracy rate of 0.84. The study also shows that Random Forest can improve accuracy compared to Decision Tree because it uses many decision trees.

Meanwhile, in this study using the XGBoost algorithm, a matrix was obtained showing that the model was able to classify 332 students as not passing and 166 students as passing. The XGBoost algorithm is very effective at recognizing patterns in student academic data based on variables like GPA, credits, attendance, average grades, and activity, making it suitable to be used as a model to predict student graduation.

METHOD

The research method used in this study is a quantitative method with a Data Mining and Machine Learning approach[4]. This method is used to build a student graduation prediction model based on historical academic data.

XGBoost builds trees sequentially with the goal of fixing the prediction errors made by the previous trees [5]. The final XGBoost model can be represented as a sum of several decision trees.

$$y= \sum_{K=1}^{K} f_k(x_i)$$

Description :
yi = prediction result
f_k = tree k
K = number of trees

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 3, June 2026, hlm 509 – 514 ISSN 2550-0201 (Online)
DOI: https://doi.org/ 10.33330/jurteksi.v12i3.4644
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

The final decision doesn't come from just one tree, but is a combination of all the boosting trees built during the training process. The diagram in this study can be illustrated as follows:

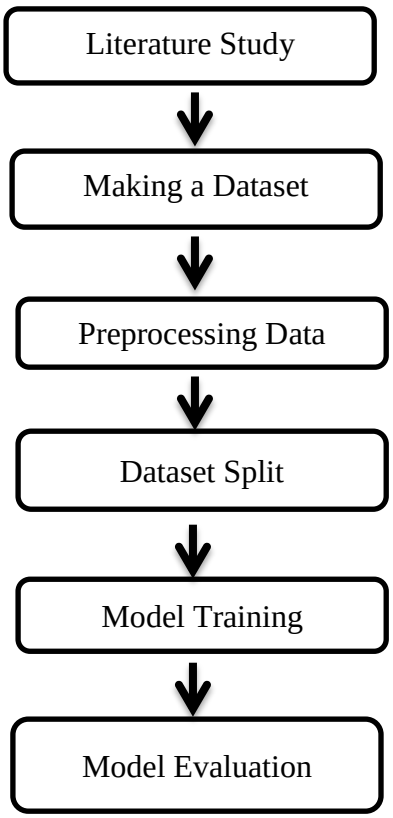

Image 1. Research Diagram

RESULTS AND DISCUSSION

Student graduation prediction is one of the applications of data mining and machine learning in the field of education. The main goal of this research is to help the campus identify students who are likely to graduate on time and those who are at risk of delayed graduation.[6]. With that prediction, educational institutions can take preventive steps like academic guidance, monitoring, and student support.

The simulation dataset used in this study consists of 500 students obtained from educational institutions. The variables used include:

Table 1. Dataset Variables

Variables	Information
IPK	Cumulative Grade Point Average
Presence	Student attendance percentage
Average score	Overall academic score
SKS	The number of credits completed
Activity	Active or nactive status

The next step is dividing the data into training data and testing data. In creating the first tree, training data is used, with the initial predicted values as follows:

Table 2. Initial Prediction

Condition	Prediction
IPK > 2.87	Active
IPK ≤ 2.87	Inactive

For the formation of the first tree, we look for the attribute with the highest gain value. Based on calculations using RapidMiner, the first tree formed is GPA because the highest gain value is GPA.

The initial prediction of students graduating is the number of students who graduate divided by the number of student samples used in this study. XGBoost's initial prediction initialization starts with the initial probability. With a simulation dataset of 500 students, with 380 students passing and 120 students failing, the calculation of the initial value is:

$$\text{Residual}=Y_{\text{aktual}} - Y_{\text{Prediksi}}$$

$$P(Y=1)=\frac{380}{500}=0{,}668$$

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 3, June 2026, hlm 509 – 514 ISSN 2550-0201 (Online)
DOI: https://doi.org/ 10.33330/jurteksi.v12i3.4644
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

$$Y_0 = \log\left(\frac{0.76}{1-0.76}\right)$$
$$Y_0 = \log(3.1667)$$
$$Y_0 = 1.152$$

Next, calculate the sigmoid probability using the following rules:

$$P_i = \frac{1}{1+e^{-1.152}} = 0{,}76$$

From the sigmoid calculation, the first iteration result is predicted to be 0.76. Calculating the gradient and Hessian, the data used is the first data point, which means y = 1 with a predicted value of 0.76, so the gradient value obtained is as follows:

$$g_i \text{ (Gradient)} = P_i - Y_i$$
$$g_i \text{(Gradient)} = 0.76 - 1 = -0.24$$

Calculate the Hessian value using the following terms:

$$h_i(\text{ hessian }) = P_i - (1 - P_i)$$
$$h_i(\text{ hessian }) = 0.76 - (0.24) = 0.1824$$

The next step is to find the best split using the xgboost model. The data choice for the split is GPA > 2.75, where the data is divided into the left node with GPA < 2.76 and the right node with GPA > 2.75. The split values can be obtained as Left Node $G_L = -40$ and $H_L = 20$, while Right Node $G_R = 40$ and $H_R = 60$.

Next, determine the gain value as the root of the tree, using the xgboost model. The XGboost model formula can be described as follows :

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} + \frac{G^2}{H + \lambda}\right] - Y$$

Terms of the data used, value $\lambda = 1$ and value $\Upsilon = 0$, then the gain value is obtained as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{40^2}{21} + \frac{40^2}{61}\right] - Y$$
$$\text{Gain} = \frac{1}{2}[76.19 + 26.23] = 51.21$$

The gain value produces a positive and large value, so the IPK split value > 2.75 is chosen as the root of the tree. After getting the highest gain value, the next step is to calculate the leaf (branch) value. The following are the provisions:

Left node $W_L = -\frac{GL}{H_L + \lambda}$

$$W_L = -\frac{-40}{21} = 1.905$$

Right node $W_L = -\frac{40}{61} = 0.656$

Next, to determine the probability value, we need a value for the new prediction. With a predicted GPA of 3.73, it goes into the right node for the new prediction, which is:

$$F_1(x) = F_0(x) + \eta W_R$$

For the grade $\eta = 0.3$

$$F_1(x) = 1.152 + 0.3\ (-0.656) = 0.955$$

So we get the probability values as follows:

$$P_i = \frac{1}{1+e^{-0.955}} = 0.722$$

It was found that the percentage of students predicted to pass is 72.2%. To get the accuracy score from this research, the researcher used the RapidMiner application. Here are the Confusion Matrix values from the accuracy classification:

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 3, June 2026, hlm 509 – 514 ISSN 2550-0201 (Online)
DOI: https://doi.org/ 10.33330/jurteksi.v12i3.4644
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

Accurasi value : 99.60%

	Not pass	Pass	Class Precision
Pred. L	332	1	99.70%
Pred. TL	1	166	99.40%
	Not pass	Pass	Class Precision
Class Recall	99.70%	99.40%	

Based on the test results using RapidMiner with the XGBoost algorithm, it can be seen as below.

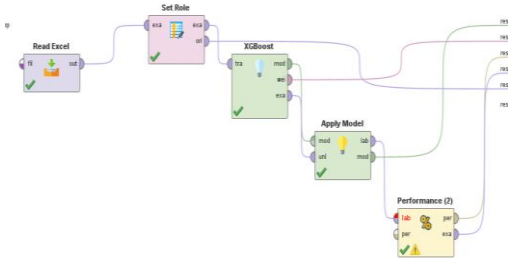

Image 2 RapidMiner Process Flow

accuracy: 99.60%

	true lulus	true tidak lulus	class precision
pred. lulus	332	1	99.70%
pred. tidak lulus	1	166	99.40%
class recall	99.70%	99.40%	

Image 3. XGboost Evaluation Results

Table View Plot View

classification_error: 0.40%

	true lulus	true tidak lulus	class precision
pred. lulus	332	1	99.70%
pred. tidak lulus	1	166	99.40%
class recall	99.70%	99.40%	

Image 4. Error classification

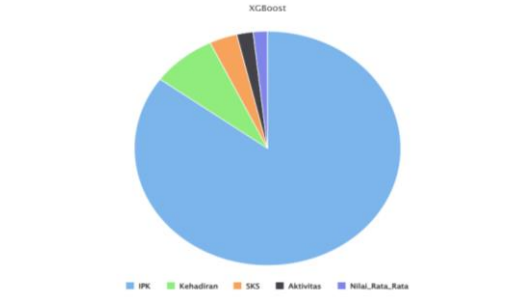

Image 4. Attribute Weight (XGBoost)

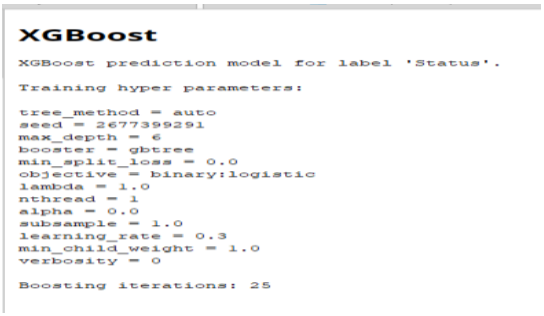

Image 5 Parameters (hyperparameters) of the XGBoostmodel

From the training process using the XGBoost algorithm in Altair AI Studio, the hyperparameter configuration obtained was tree_method = auto, max_depth = 6, booster = gbtree, objective = binary:logistic, learning_rate = 0.3, lambda = 1, alpha = 0, subsample = 1, and the number of boosting iterations was 25. This configuration was used to create a classification model that can predict students' graduation status based on GPA, number of credits, attendance, and average grades.

The use of the binary:logistic objective is suitable because the study only has two classes, namely Pass and Fail, while L2 regularization (lambda = 1) helps reduce the risk of overfitting. The model that was made was then evaluated using a confusion matrix and got an accuracy of 99.60%, showing that this hyperparameter configuration can produce very good classification performance on the simulated dataset.

BIBLIOGRAPHY

[1] N. A. Sivi *et al.*, "Prediksi Kelulusan Mahasiswa UNU Lampung Menggunakan Algoritma Decision Tree Berbasis Data Akademik Menggunakan Rapidminer," 2026.

[2] D. A. Punkastyo, F. Septian, and A. Syaripudin, "Implementasi Data Mining Menggunakan Algoritma Naïve Bayes Untuk Prediksi Kelulusan Siswa," vol. 5, no. 1, pp. 24–35, 2024.

[3] S. Junaidi, R. V. Anggela, and D. Kariman, "JOURNAL OF APPLIED COMPUTER SCIENCE AND TECHNOLOGY (JACOST) Klasifikasi Metode Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa dengan Algoritma Naïve Bayes , Random Forest , Support Vector Machine (SVM) dan Artificial Neural Nerwork (ANN)," vol. 5, no. 1, pp. 109–119, 2024.

[4] R. Mitchell and E. Frank, "Accelerating the XGBoost algorithm using GPU computing," 2017, doi: 10.7717/peerj-cs.127.

[5] E. Science, "Application of XGBoost algorithm in hourly PM2 . 5 concentration prediction Application of XGBoost algorithm in hourly PM2 . 5 concentration prediction."

[6] N. Y. Arifin, P. Studi, T. Informatika, U. I. Sina, K. Riau, and D. Simulasi, "Studi Metodologis Optimasi Hyperparameter XGBoost Menggunakan Bayesian Optimization untuk Prediksi Risiko Stunting Berbasis Dataset Simulasi," vol. 3, no. 01, pp. 41–48, 2026.