Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

June 2026. Vol. 5. No. 3; e-ISSN: 2808-4519

Early Warning System for Detecting Student Dropouts Using the Random Forest Algorithm at SMKS Alhuda

Muhammad Khoerulrijjal Salsabila Januarta[1]*, Nuk Ghurroh Setyoningrum[2]

[1,2] *Cipasung University, Tasikmalaya*
rijalsaja73@gmail.com [1*], *nuke@uncip.ac.id* [2]

Abstract

The dropout rate in vocational high schools poses a serious challenge that requires an objective early-detection system. This study aims to optimize a model for predicting student dropout risk by utilizing a supervised learning approach. The study uses multivariate data covering student demographic attributes, academic achievement, behavior, and financial history. The Random Forest algorithm was implemented to classify student risk levels into Safe, Caution, and Danger categories to support preventive decision-making. Model performance testing using a confusion matrix showed an accuracy rate of 99%, with a recall of 100% in the High-Risk category, demonstrating the algorithm's effectiveness in accurately identifying high-risk students. These findings contribute to the development of more precise early detection methods in educational settings.

Keywords: *Dropout Prediction, Early Warning, Machine Learning, Random Forest, Vocational Education*

1. Introduction

Economic difficulties and a lack of discipline are often the main obstacles preventing vocational high school (SMK) students from completing their education on time. Unstable family finances have proven to be a dominant factor driving students out of the education system , with the inability to cover school fees being the primary reason students drop out[1] . This economic pressure is often exacerbated by low parental education levels and a lack of encouragement from the surrounding community, causing students to lose motivation to continue their studies[2] . At the school level, the process of identifying students at risk of dropping out is generally still conducted manually and reactively, meaning at-risk students are often only detected after entering a critical phase, leaving very limited room for preventive intervention[3] .

Machine learning-based approaches have proven effective in building predictive models for dropout prevention. Student attendance and absence variables are among the most significant predictors in determining the probability of dropping out[4] , while financial conditions such as a history of administrative payments and parental economic capacity also make a significant contribution to students' continued enrollment[5] . Research based on academic data shows that a combination of variables such as GPA, attendance rates, and course load can produce predictive models with high accuracy in classifying at-risk students[6] .

Fig. 1: Enrollment trends and dropout rates, 2021–2025

Based on Figure 1, the dropout rate shows a consistent increase each year, starting from 11.0% in 2021 and reaching 15.3% in 2025. This trend confirms that without a structured detection system, the risk of dropout will continue to rise. The absence variable contributes 28% and financial history 35% to the accuracy of the prediction model, making the combination of these two parameters crucial in strengthening risk classification in this study.

When processing heterogeneous multivariate datasets, the Random Forest algorithm demonstrates high stability and accuracy in classifying complex tabular data, consistently outperforming other algorithms across various testing scenarios[7] . The implementation of an early warning system that integrates financial and academic variables has proven reliable in detecting potential dropouts early on and supporting more planned preventive interventions in educational institutions[8] . Support from digitized information systems is also a key factor, as the centralized integration of academic and administrative data has been shown to improve decision-making efficiency by school administrators[9] .

Although many similar studies have been conducted, the majority of previous studies have focused on higher education institutions with complex variables, or have relied solely on academic aspects. Predictive models specifically designed for vocational high schools that integrate financial aspects (tuition arrears) with daily performance data remain extremely rare[10] . Most existing models are also solely focused on overall accuracy without generating a risk-level mapping specific to each student.
Based on these issues, this study focuses on optimizing a dropout risk classification model using the Random Forest , by incorporating multivariate parameters that include demographic, academic, disciplinary, and financial history data from 1,000 students. This study aims to produce a precise prediction model based on a confusion matrix as a strategic decision-making tool for schools, by objectively mapping risks into three categories: Safe, Caution, and Danger.

2. Research Methodology

The methodology used in this study was systematically designed to develop a model for classifying student dropout risk. The research workflow was adapted from a standard data science experimental framework that covers phases ranging from data acquisition to evaluation, as illustrated in Figure 2.

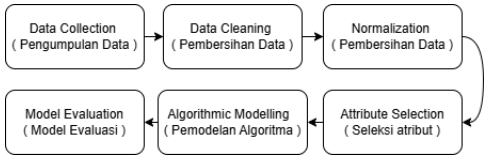

Fig.2: Research Phases

Figure 2 illustrates a cyclical and iterative research workflow. To achieve the research objectives, the experimental stages were conducted as follows:

1. Data Collection. The initial phase involved acquiring historical data from 1,000 students at SMKS Al-Huda, comprising 28 multivariate attributes. This tabular data includes multivariate variables grouped into four main parameters: demographics (socioeconomic background), academics (learning performance), discipline (attendance records with a 28% accuracy contribution), and financial history (tuition administration with a correlation of up to 35%).
2. Data Cleaning. Once the data has been collected, a cleaning process is performed to ensure the quality of the model inputs. This stage involves handling missing values through statistical imputation, removing irrelevant data, and deduplicating the data to maintain data integrity and prevent bias in the prediction results[11] .
3. Data Normalization. The cleaned data then undergoes a normalization process using Min-Max Scaling or Standardization techniques. This stage aims to align the scale of numerical features to prevent the dominance of certain variables with large absolute values during algorithm processing, thereby making computational performance more stable[12] .
4. Attribute Selection. This step is performed to select the multivariate parameters that have the greatest influence on risk estimates. Using correlation or feature importance techniques, researchers retain the most significant attributes and discard redundant ones to improve the model's efficiency and accuracy[13] .
5. Algorithm Modeling. In this stage, the Random Forest algorithm is applied due to its high stability in handling multivariate tabular datasets and its ability to minimize overfitting[14] . The algorithm builds an ensemble of decision trees during the training phase, where the final decision is made through a majority voting mechanism to classify student risk into the categories Safe, Caution, and Danger.
6. Model Evaluation. The final step is to evaluate the model's performance using a confusion matrix. This tool calculates overall accuracy and maps the model's predictions against the actual classes to ensure that the resulting early warning system has high precision before it is implemented by schools for preventive guidance[15] .

   To measure the model's performance, the following evaluation formula is used:

   a. Accuracy:

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

   Explanation:
   TP: True Positive
   FP: False Positive
   TN: True Negative
   FN: False Negative

   b. Precision:

$$\text{Prec (Precision)} = \frac{\text{TP}}{\text{TP+FP}}$$

   Notes:
   TP: True Positive
   FP: False Positive

c. Sensitivity/Recall:

$$\text{Rec (Recall)} = \frac{\text{TP}}{\text{TP+FN}}$$

Notes:
TP: True Positive
FN: False Negative

The results of this evaluation determine whether the developed model meets operational standards, so that the school can implement targeted interventions for students in the high-risk category.

3. Results And Discussion

This section presents the results of the *dropout* risk classification experiment using the *Random Forest* algorithm. The main focus of the presentation is on the effectiveness of the predictive model in identifying multivariate data patterns of students based on academic, disciplinary, and financial parameters.

3.1 Data Transformation and Preparation

The research *dataset* consists of 1,000 student records with raw multivariate attributes. The initial phase of the research involved *data* cleaning, during which personal identifiers—such as student names—and other attributes not included in the 28 primary research attributes were removed to ensure the model's objectivity and compliance with data privacy regulations. The raw data before preprocessing is presented in Table 1

Table 1: s of Data Before Preprocessing

Student Name	Gender	School_Origin	Residence_Status	Father's_Occupation	Scholarship_Status	…
Wahyu Sanjaya	L	Private	Brother	Merchant	No	…
Hafiz Imron	L	Private	Boarding house	Merchant	Yes	…
Hamdan Dwi	L	Private	Brother	Self-employed	No	…
Rendi Halim	L	Private	Brother	Brother	No	…

Based on Table 1, *the dataset's* condition prior to cleaning is evident; it still contains personal identification data as well as attributes irrelevant to the analysis's needs. The selection process was conducted to filter out these attributes so that *the dataset* would be more focused on parameters influencing the risk of dropping out of school.

After undergoing data cleaning and feature selection, *the dataset* was further processed through a data transformation stage. Since *the dataset* contained categorical variables (such as economic status), *the One-Hot Encoding* technique was applied to transform the textual data into a binary numerical representation so that it could be processed mathematically by the *Random Forest* algorithm[16] . The data after preprocessing is presented in Table 2

Table 2: Data After Preprocessing

Total_Subjects	Math_Score	Indonesian_Score	BIng_Score	Basic_Product_Score	…
10	78	82	84	92	…
12	83	68	81	64	…
15	78	84	82	81	…
10	85	84	77	87	…

Table 3.2 shows the final form of *the dataset* after it has undergone *preprocessing* and *encoding*. All categorical variables have been converted to numerical format, enabling the algorithm to perform risk calculations and classifications objectively and with precision.

3.2. Model Training and Experiment Stability

The experiment used a *hold-out* data split scheme with an 80% training data and 20% test data ratio. The model was built with a configuration of 100 *decision trees* to ensure stability and minimize the risk *of overfitting*. To monitor performance consistency during the training process, a learning curve visualization was used.

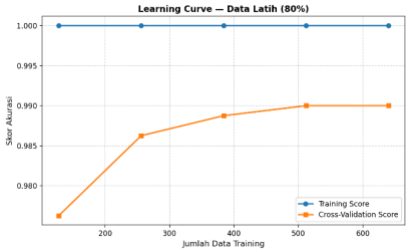

Fig.3: 's *Learning Curve*

This graph shows how the model's performance improves as the amount of training data increases. The convergence of the training score and cross-validation score curves confirms the algorithm's stability in handling multivariate *datasets* and indicates that the model is not overfitting[17] .

3.3. Classification Performance Evaluation

A formal test was conducted on 200 test data points to measure the model's precision in mapping the risk categories Safe, Caution, and Danger. Based on this test, the model demonstrated highly precise performance. The "Danger" category was successfully identified with a perfect *recall* value (1.00), meaning no high-risk students were missed by the model. Performance details for each class are presented in Table 3 below.

Table 3: Model Evaluation Results

--- Model Evaluation Results ---				
	precision	**recall**	**F1-score**	**support**
Safe	1.00	0.99	1.00	119
Danger	1.00	1.00	1.00	31
Caution	0.98	1.00	0.99	50
Accuracy			0.99	200
Macro Average	0.99	1.00	1.00	200
Weighted Average	1.00	0.99	1.00	200

The evaluation results above confirm that this model is highly reliable, with an overall accuracy of 99%. The high consistency of *precision* and recall scores across all categories indicates that the model can serve as an accurate strategic decision-making tool for schools in implementing preventive interventions.

These numerical results are further reinforced by the visualization in the following figure:

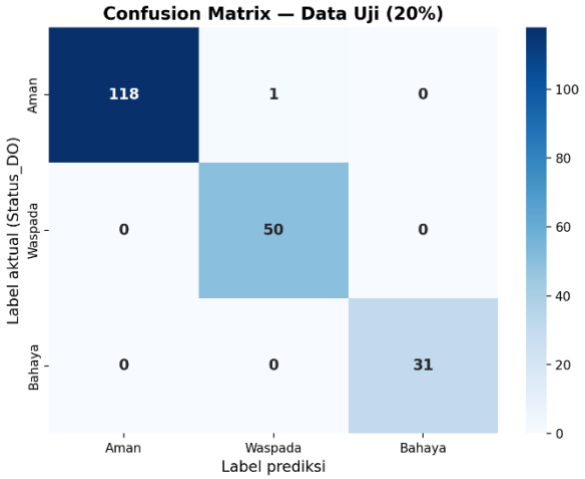

Fig.4: Confusion Matrix

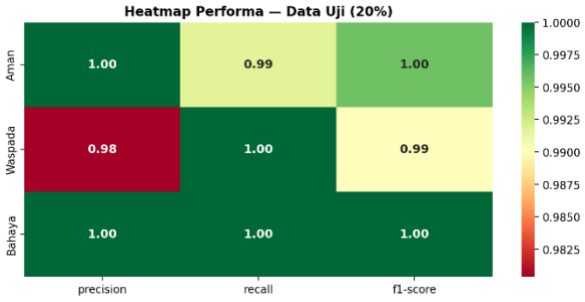

Fig.5: Performance Heatmap

Based on Figure 3.2 and Figure 3.3, *the Confusion Matrix* provides quantitative details regarding the number of correct and incorrect predictions for each class, while *the heatmap* offers an intuitive visualization of the accuracy distribution. The dominance of color intensity along the main diagonal of *the heatmap* clearly validates that the model has a very high level of accuracy, with minimal misclassification between classes.

The model's ability to distinguish between risk classes is also measured using the ROC curve.

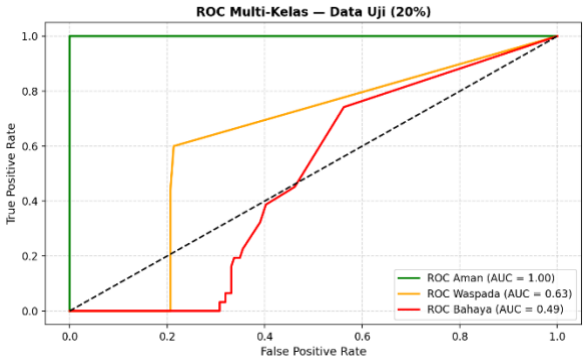

Fig.6: ROC (*Receiver Operating Characteristic*)

As shown in Figure 3.4, the Area Under the Curve (AUC) value, which is close to 1.00, demonstrates that the *Random Forest* algorithm is highly effective in distinguishing high-risk student profiles from those in the safe category, with a high *True Positive Rate* ([18]).

3.4. Attribute Significance Analysis and Decision Logic

It is important to understand which variables most influence risk estimation through *Feature Importance* analysis.

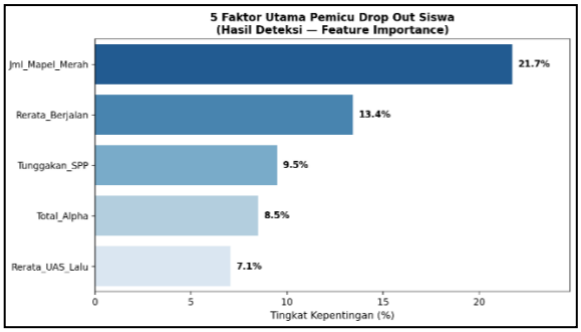

Fig.7: Feature Importance Analysis

Based on this analysis, it was revealed that academic performance and financial arrears are the most dominant predictors with significant influence weights[19] . This proves that economic factors have a strong correlation with students' ability to continue their studies at private schools.

Finally, to provide transparency regarding the algorithm's internal mechanisms, a sample of one decision tree from *the ensemble forest* was selected.

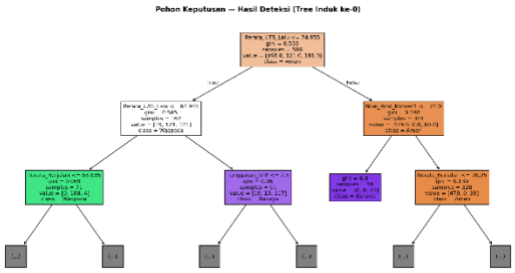

Fig.8: Visualization of *the Decision* Tree Structure

Based on Figure 3.6, this visualization illustrates the logical flow of attribute branching (such as attendance rates and tuition payment status) used by the model to automatically determine students' risk status, thereby helping guidance counselors understand the basis for the model's decision-making[20] .

4. Conclusion

Based on the test results conducted by the researcher, it can be concluded

1. That the *Random Forest* model with a configuration of 100 decision trees is capable of classifying student dropout risk consistently. This is evidenced by the consistent performance on the learning curve, which shows no signs of overfitting.
2. The integration of financial data (outstanding tuition fees) proved to be a significant determining variable. Based on the feature importance analysis, this factor has a strong correlation with the accuracy of dropout risk predictions, surpassing the influence of academic and disciplinary variables.
3. Model evaluation results using *a confusion matrix* show a high level of accuracy in classifying students into the Safe, Caution, and Danger categories, with an AUC value approaching 1.00. This indicates that the algorithm has excellent class separation capabilities.

5. Recommendations

The recommendations the author can offer are as follows:

1. The dropout risk classification model that has been developed can still be further refined as artificial intelligence technology advances. Given the complexity of the factors causing school dropout, this system needs to be integrated with a more comprehensive monitoring module so that the information generated can cover students' psychosocial aspects in greater depth.
2. For future researchers, it is hoped that they can expand *the dataset* by including variables related to the school environment, such as family support and social conditions, to improve the model's accuracy and sensitivity in identifying student risk profiles that are not captured by internal school data.
3. To address potential data *imbalance* that may occur in the future, it is recommended that future researchers implement more advanced data balancing techniques, such as *SMOTE (Synthetic Minority Over-sampling Technique*) or *ADASYN*, to ensure the model remains consistent in identifying risk classes with limited sample sizes.

References

[1] N. A. Vita, "Analisis Faktor Penyebab Meningkatnya Angka Putus Sekolah di Indonesia pada Tahun 2022," *J. Pendidik. Sultan Agung*, vol. Vol. 03 No, no. 005, p. 177, 2023.

[2] S. Frisnoiry, "Analisis Faktor Penyebab Anak Putus Sekolah," *J. Cendekia Ilm.*, vol. 3, no. 5, pp. 2480–2492, 2024.

[3] A. Sholihin Fauzan, A. Irma Purnama Sari, and I. Ali, "Analisis Perbandingan Algoritma Decisioin Tree Dan Naïve Untuk Mengevaluasi Prestasi Belajar Siswa," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 741–747, 2024, doi: 10.36040/jati.v8i1.8403.

[4] A. Algiffary and T. Sutabri, "Indonesian Journal of Computer Science," *Indones. J. Comput. Sci.*, vol. 12, no. 2, pp. 284–301, 2023, [Online]. Available: http://ijcs.stmikindonesia.ac.id/ijcs/index.php/ijcs/article/view/3135

[5] L. G. R. Putra, D. D. Prasetya, and M. Mayadi, "Student Dropout Prediction Using Random Forest and XGBoost Method," *INTENSIF J. Ilm. Penelit. dan Penerapan Teknol. Sist. Inf.*, vol. 9, no. 1, pp. 147–157, 2025, doi: 10.29407/intensif.v9i1.21191.

[6] P. S. Saputra, "Analisis Prediktif Dropout Mahasiswa Berdasarkan Kinerja Akademik Semester Awal Menggunakan Machine Learning," *J. Ris. dan Apl. Mhs. Inform.*, vol. 07, no. 01, pp. 164–171, 2026.

[7] S. Kasus, K. Putusan, and M. Konstitusi, "Jurnal Sains Informatika Terapan (JSIT) Jurnal Sains Informatika Terapan (JSIT)," pp. 187–201, 2025.

[8] F. A. Putra, S. Mirajdandi, B. Okmarizal, and S. Mulyanda, "Prediksi Dropout Mahasiswa : Early-Warning Berbasis Enrollment dengan Machine Learning," vol. 15, no. 3, pp. 465–473, 2025.

[9] M. B. Firdaus, H. Rakhmawati, R. Fauziyah, and H. C. Prakoso, "PENGEMBANGAN SISTEM INFORMASI MANAJEMEN SEKOLAH BERBASIS WEBSITE DI SD NEGERI LANGKAP 01," vol. 13, no. 1, 2026.

[10] Filan Firmansyah, Saputra Dwi Nurchaya, and Zuhana Realita Alfy, "Perbandingan Model Pembelajaran Mesin Berbasis Smote Meningkatkan Identifikasi Siswa Berisiko di Sekolah Menengah Pertama," *JSiI (Jurnal Sist. Informasi)*, vol. 11, no. 2, pp. 1–6, 2024, doi: 10.30656/jsii.v11i2.9065.

[11] L. Hakim, A. Sobri, L. Sunardi, and D. Nurdiansyah, "Prediksi penyakit jantung berbasis mesin learning dengan menggunakan metode k-nn," *J. Digit. Teknol. Inf.*, vol. 7, no. 2, p. 14, 2025, doi: 10.32502/digital.v7i2.9429.

[12] I. Permana, "The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation," *Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022, [Online]. Available: https://media.neliti.com/media/publications/485639-pengaruh-normalisasi-data-terhadap-perfo-e19e3a00.pdf

[13] E. Novianto, S. Suhirman, and D. Prasetyo, "Perbandingan Metode Klasifikasi Random Forest Dan Support Vector Machine Dalam Memprediksi Capaian Studi Mahasiswa," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 4, pp. 1821–1833, 2024, doi: 10.29100/jipi.v9i4.5423.

[14] M. Mahendra Alvanof and R. Kesuma Dinata, "Penerapan Algoritma Random Forest dalam Deteksi dan Klasifikasi Ransomware," *J. Elektron. dan Teknol. Inf.*, vol. 5, no. 2, pp. 2721–9380, 2024.

[15] T. H. Pinem and Z. P. Putra, "Evaluasi Kinerja Algoritma Klasifikasi Deep Learning dalam Prediksi Diabetes," *J. Ilm. FIFO*, vol. 17, no. 1, p. 17, 2025, doi: 10.22441/fifo.2025.v17i1.003.

[16] C. Herdian, A. Kamila, F. F. Tampinongkol, A. S. Kembau, and I. G. A. M. Budidarma, "One-hot encoding feature engineering untuk label-based data studi kasus prediksi harga mobil bekas," *Inf. Interaktif J. Inform. dan Teknol. Inf.*, vol. 9, no. 1, pp. 10–16, 2024, doi: 10.37159/jii.v9i1.41.

[17] G. A. M. Ashfania, T. Prahasto, A. Widodo, and T. Warsokusumo, "Penggunaan Algoritma Random Forest untuk Klasifikasi berbasis Kinerja Efisiensi Energi pada Sistem Pembangkit Daya," *Rotasi*, vol. 24, no. 3, pp. 14–21, 2023.

[18] Z. A. Dwiyanti and C. Prianto, "Prediksi Cuaca Kota Jakarta Menggunakan Metode Random Forest," *J. Tekno Insentif*, vol. 17, no. 2, pp. 127–137, 2023, doi: 10.36787/jti.v17i2.1136.

[19] H. M. Nawawi, A. B. Hikmah, A. Mustopa, and G. Wijaya, "Model Klasifikasi Machine Learning untuk Prediksi Ketepatan Penempatan Karir," vol. 14, no. 1, pp. 13–25, 2024.

[20] M. Z. Ramadhany *et al.*, "KLASIFIKASI MAHASISWA BERPOTENSI DROP OUT (DO) MENGGUNAKAN ALGORITMA RANDOM FOREST," vol. 10, no. 2, pp. 3543–3548, 2026.