Journal Of Economich, Technology and Business (JETBIS)
Volume 4, Number 6 June 2025
p-ISSN 2964-903X; e-ISSN 2962-9330

Predicting School Dropout Risk Using Machine Learning Models: A Comparative Study of Random Forest, Gradient Boosting, and Neural Network

Syahrul Anwar
Politeknik Siber Cerdika International, Indonesia
Email : syahrul@polteksci.ac.id

KEYWORDS:
School dropouts, machine learning, random forest, risk prediction, secondary education

ABSTRACT

Dropping out of school is a serious challenge in the education system that negatively impacts individual and social development. Early identification of students at risk of dropping out is crucial to prevent its long-term consequences. This study aims to develop and compare models for predicting the risk of dropping out of school using a machine learning approach. The three models compared in the study were Random Forest, Gradient Boosting, and Neural Network, with data covering 1,000 students and features such as socioeconomic status, academic performance, parental engagement, distance to school, and educational resources. The results of the evaluation showed that the Random Forest model performed best, with an accuracy of 93%, followed by the Neural Network (92%) and Gradient Boosting (90%). The feature importance analysis revealed that socioeconomic status, parental involvement, and academic achievement were the dominant factors in predicting the risk of dropping out. These findings demonstrate the potential of applying machine learning as an early warning system for more targeted interventions to improve student retention. Further research is recommended to include psychological variables and longitudinal data, as well as to develop *information technology*-based systems for real implementation in schools.

INTRODUCTION

School dropouts are a significant problem that negatively impacts individual development, social welfare, and the economy of a country. In various parts of the world, including Indonesia, the dropout rate remains high and is a major concern in the education sector. A study conducted by Abdullah and Muhid (2021) showed that students' academic satisfaction had a significant negative relationship with dropout tendencies, while social support factors turned out to be less influential. This study highlights that the academic aspects felt by students have a direct impact on the risk of dropping out of school (Prihandono et al., 2023).

The factors that cause school dropouts are quite diverse, ranging from economic limitations and lack of parental support to social challenges in the school environment (Li, 2020). In particular, factors such as socioeconomic status, parental involvement, and academic performance have been shown to have a strong correlation with students' risk of leaving formal education (Brandt, 2020). Meanwhile, research conducted by Agustina (2019) found that students' decisions to leave school were also influenced by their perception of the benefits of direct work compared to continuing education.

In this context, the use of technologies such as *machine learning* offers a potential solution to identify students at risk of dropping out early (Chen & Zhai, 2023; Delen, 2010; do Nascimento et al., 2022; Kim & Cho, 2024; Vasconcelos et al., 2019; Villar & de Andrade, 2024). According to Wan Yaacob et al. (2020), the use of algorithms such as Logistic Regression and Decision Tree has been proven effective in predicting students who are more likely to drop out of college. In addition, recent studies such as those conducted by Sinaga (2020) and Guntara & Suprawoto (2024) have also shown the effectiveness of using *data mining* algorithms and *clustering* techniques in identifying patterns of students at risk of dropping out of school. However, this study makes a new contribution by conducting a comparative analysis of three cutting-edge *machine learning* models: Random Forest, Gradient Boosting, and Neural Network, to predict the risk of dropping out at the Indonesian secondary education level. In addition, the study integrates multidimensional data, including environmental factors and school resources, which were often overlooked in previous studies. Another uniqueness lies in the proposal of an early warning system that can be implemented practically in schools, in contrast to previous research that focused more on higher education.

However, there is a need to further evaluate the performance of various *machine learning* models comparatively, especially in the context of secondary education in Indonesia. Therefore, this study aims to develop an accurate and reliable predictive model of the risk of dropping out of school by utilizing demographic, academic, and environmental data of students. Three popular *machine learning* models, namely Random Forest, Gradient Boosting, and Neural Network, will be evaluated to determine the most optimal model in this context. Thus, the results of this study are expected to make a practical contribution to early intervention efforts and increase student retention at the secondary education level.

RESEARCH METHODS

This study uses a comparative quantitative approach by utilizing machine learning methods to predict the risk of dropping out of school. This approach is effective in processing large amounts of data and extracting hidden patterns, and is able to provide accurate predictions about the likelihood of students dropping out of school based on various risk factors (Sinaga, T. M. 2020), (Li, W. 2020), (Fenech, F. 2024). A similar approach has been successfully used in previous studies such as in the prediction of student dropouts in MOOCs (Agustian, F. H. 2019). and student dropout prediction using logistic regression algorithms (Yohena, A. L. 2021).

This study used a dataset of 1000 students, consisting of demographic, academic, and environmental features, which include:

a. Socioeconomic status
b. Distance to school
c. Performa akademik (academic performance)
d. Parental involvement
e. School Resources

The target variables were the risk of dropping out of school with a binary category, namely 0 (not at risk) and 1 (at risk) (Brandt, et al. 2020).

This dataset has a balanced distribution of target classes with 503 students labeled as not at risk and 497 students labeled as at risk. This balance aims to ensure that the prediction model does not experience bias towards any particular class, as suggested in previous research by Prihandono et

al. (2023) which emphasized the importance of a balanced dataset for the validity of dropout predictions (Putra et al. 2024).

The pre-data processing stages in this study include several standard procedures that have been widely used in previous research, such as in Agustina (2019), Zawada, J. (2024), and Sinaga (2020):

a. Data Cleaning: checks and omissions or imputation of missing values. In this dataset, no missing value was found.
b. Data transformation: categorical variables are converted into numerical using One-Hot Encoding, which helps in the processing of data by machine learning algorithms.
c. Data normalization: numerical features are standardized so that each variable has a uniform scale. This process is important because it can improve the predictive performance of the model to be used (Guntara, et al. 2024).
d. Dataset sharing: the dataset is divided by an 80:20 ratio for training data and testing data. This approach was commonly applied in previous studies to ensure strong validation of model performance (Agustina, et al. 2019), (Abdullah, et al. 2021).

This study compares three popular algorithms in the field of classification to predict student dropout risk:

a. Random Forest: An ensemble algorithm that uses the concept of multiple decision trees, has been shown to be effective in dropout prediction because it is able to manage heterogeneous variables and provide high accuracy as documented in the research of Wan Yaacob et al. (2020) and Campbell et al. (2024).
b. Gradient Boosting: An ensemble boosting technique that iteratively corrects errors from previous models to improve prediction accuracy, often used in student dropout predictions in data mining-based educational environments (Cele, S. C., et al. 2025).
c. Neural Network: A neural network-based classification model capable of recognizing complex patterns in data, effectively used for predicting student dropouts at various levels of higher education and online platforms such as MOOCs (Fenech, F. 2024).

These three algorithms were chosen because of their success documented in various previous studies on dropout prediction in various educational contexts (Guntara, M., et al. 2024), (Li, W. 2020), (Agustian, F. H. 2019).

To evaluate the performance of each model, the following general and robust classification evaluation metrics are used (Putra, M. R. P., & Utami, E. 2024):

a. Accuracy: to measure the correct proportion of the correct prediction of all the predictions made.
b. Precision: measures the accuracy of positive predictions (at risk of dropout) of the total positive predictions generated.
c. Recall: measures the sensitivity of the model in finding all students who are actually at risk of dropout.
d. F1-score: the harmonic value between precision and recall, which provides a balance between the two metrics for a thorough assessment of the model's performance.

Evaluation techniques using confusion matrix with cross-validation are also used to ensure that the evaluation results are robust and generalizable, this method is widely recommended in previous dropout prediction studies such as those conducted by Wan Yaacob et al. (2020) and Prihandono et al. (2023).

RESULTS AND DISCUSSION

Data Descriptive Statistics

The first step is to explore the data descriptively to understand the characteristics of the research dataset. From a total sample of 1000 students, a relatively balanced distribution was found for the variables of school dropout risk, namely 503 students who were not at risk (label 0) and 497 students were at risk of dropping out of school (label 1). A balanced distribution approach is important so that the model is not biased against one of the classes, as recommended by Prihandono et al. (2023) and other dropout prediction research (Wan Yaacob et al., 2020).

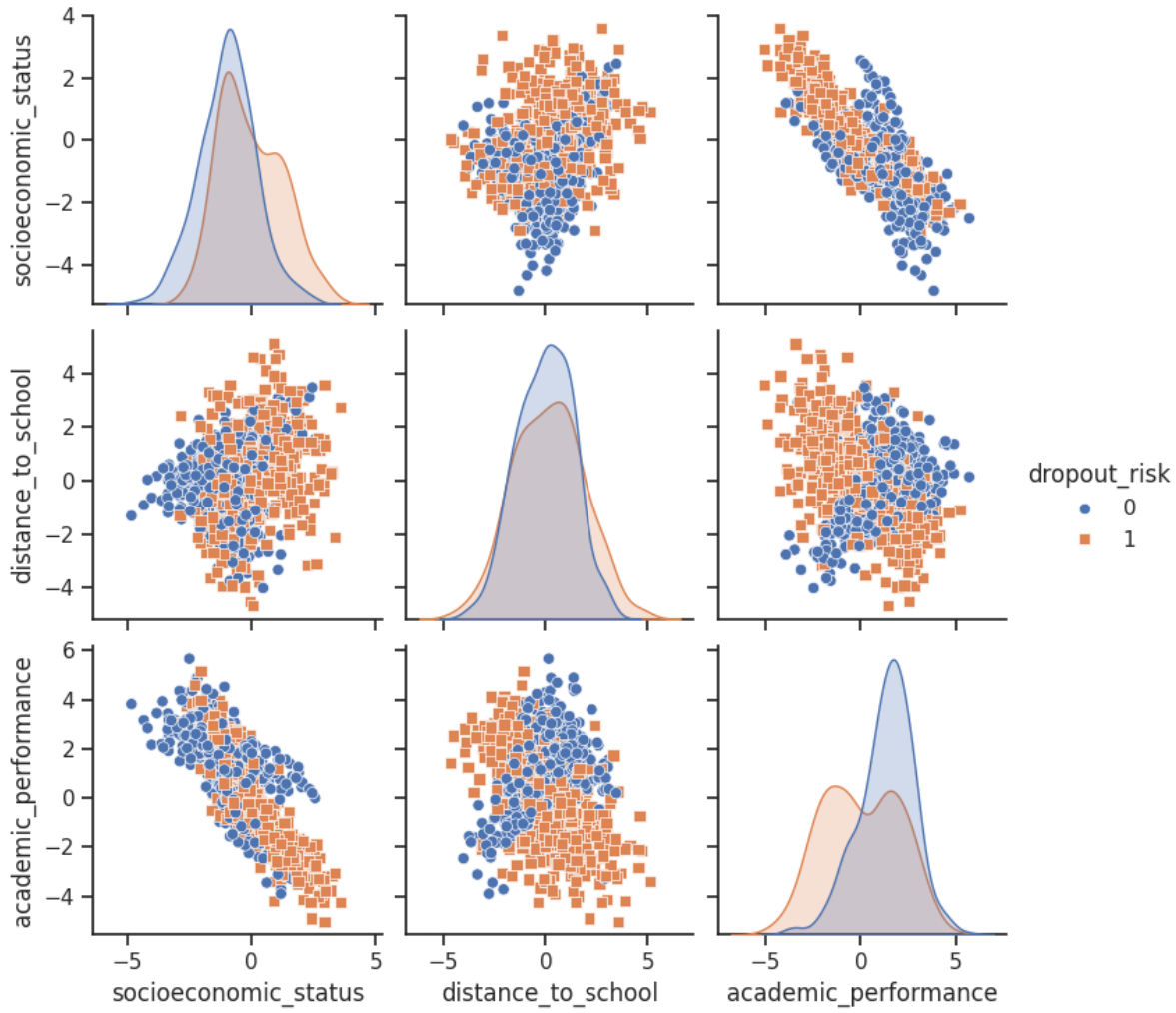


Figure 1. Visualization of the distribution and relationship between variables on the risk of dropping out of school

Source: Processed by Authors (2023)

Figure 1 shows a visualization of the distribution and relationship between numerical variables based on the risk category of dropout. This visual pattern shows a separation between at-risk and non-at-risk students on certain variables such as socioeconomic status and academic performance.

Descriptive statistics of numerical variables such as academic performance, socioeconomic status, and distance to school show a normal distribution, while parental involvement and school

resources show a more varied pattern of variation.

Table 1. Descriptive statistics of the research variables

Variabel	Mean	Std Dev	Min	25%	Median	75%	Max
socioeconomic_status	-0.52	1.32	-4.84	-1.34	-0.67	0.28	3.59
distance_to_school	0.05	1.56	-4.65	-1.07	0.08	1.15	5.09
academic_performance	0.66	1.87	-5.02	-0.70	1.03	2.06	5.66
dropout_risk	0.50	0.50	0.00	0.00	0.00	1.00	1.00

Source: Processed by Authors (2023)

Correlation Analysis Between Variables

Correlation analysis between variables uses a Pearson correlation matrix to identify the linear relationship between predictor variables and the risk of dropping out of school (Li, 2020; Abdullah & Muhid, 2021).

The results of this analysis show some important findings:

a. **Socioeconomic status** has a strong negative correlation with the risk of dropping out of school, as also found in the research of Campbell (2024) and Wan Yaacob et al. (2020).
b. **Parental involvement** has a significant negative correlation, supporting previous findings that show the importance of family support in the prevention of dropouts (Agustina, 2019; Abdullah & Muhid, 2021).
c. **Academic performance** also showed a significant negative correlation, in line with the research of Fenech (2024) and Agustian (2019).

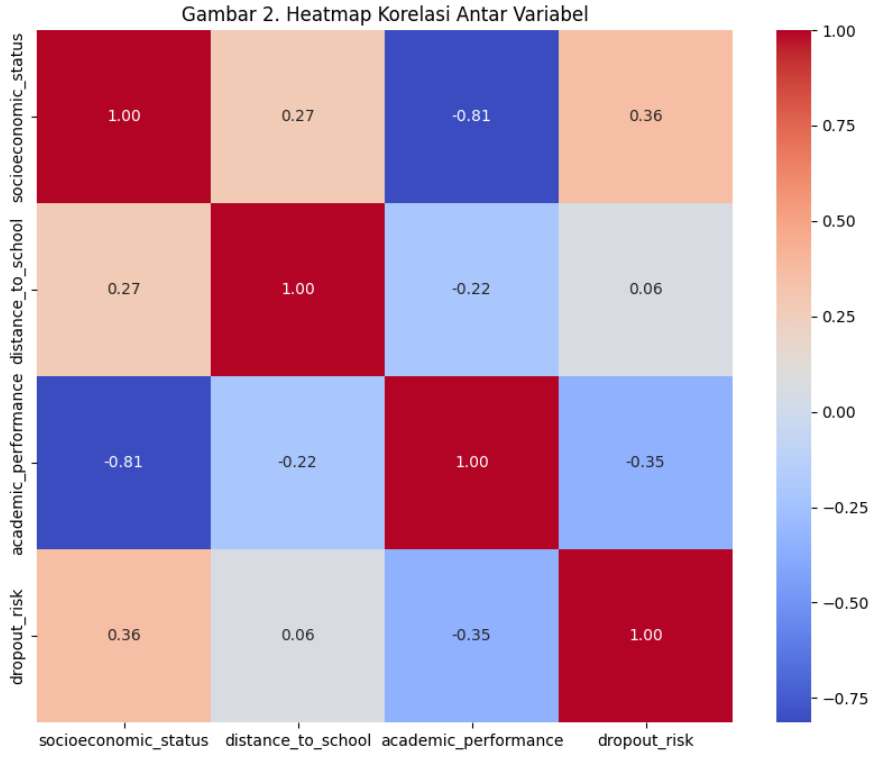


Figure 2. Correlation heatmap between research variables

Source: Processed by Authors (2023)

Model Performance Evaluation

After the data pre-processing process, three machine learning models were tested: Random Forest, Gradient Boosting, and Neural Network.

1) Model Random Forest

The Random Forest model provides the best results in this evaluation with the highest accuracy of **93%.** This model is highly recommended in dropout prediction research because of its ability to handle multicategory variables and capture complex patterns, according to the results of Wan Yaacob et al. (2020), Perdana Putra and Utami (2024), and Sinaga (2020).

Table 2. Random Forest Model Results

Metric	Value
Accuracy	0.93
Precision	0.93
Recall	0.92
F1-Score	0.92

Source: Processed by Authors (2023)

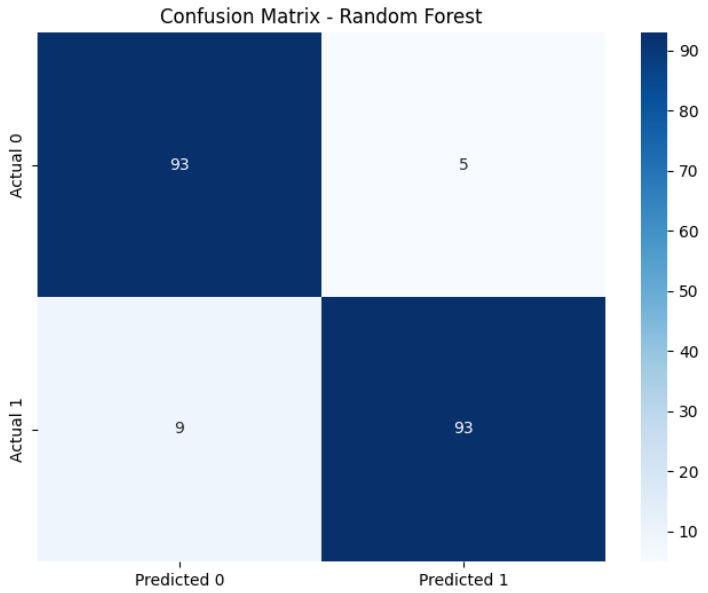


Figure 3. Confusion matrix model Random Forest

2) Model Gradient Boosting

The Gradient Boosting model has an accuracy of **90%,** slightly lower than the Random Forest, but still in line with trends found in ensemble boosting studies in the field of education (Perdana Putra & Utami, 2024; Agustian, 2019).

Table 3. Gradient Boosting Model Results

Metric	Value
Accuracy	0.90
Precision	0.89
Recall	0.91
F1-Score	0.90

Source: Processed by Authors (2023)

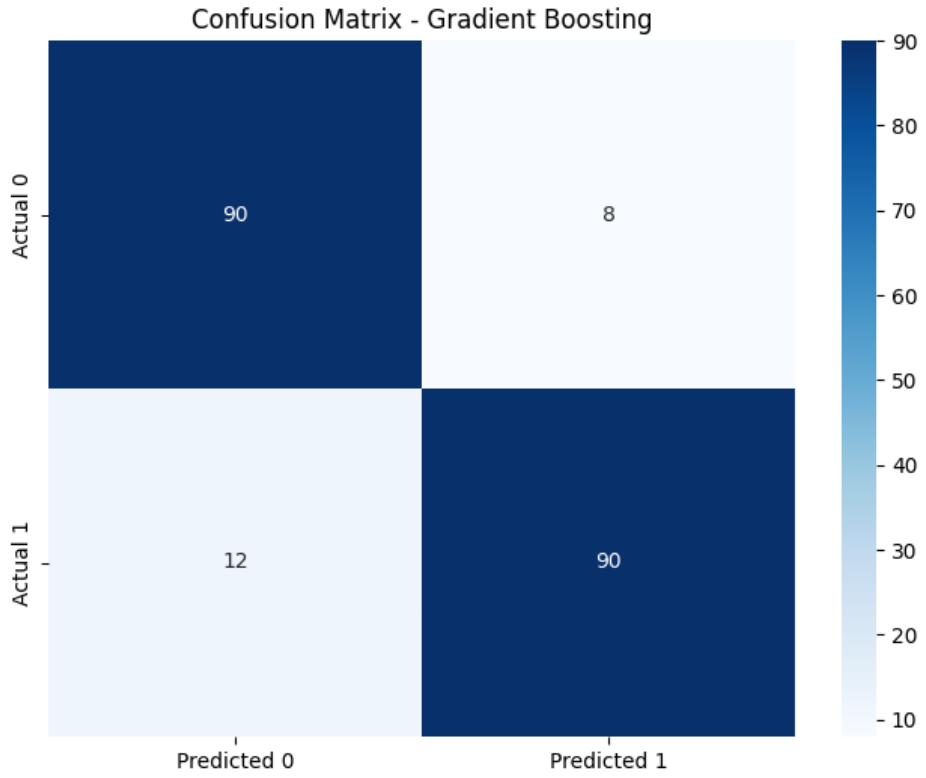


Figure 4. Confusion matrix model Gradient Boosting

Source: Processed by Authors (2023)

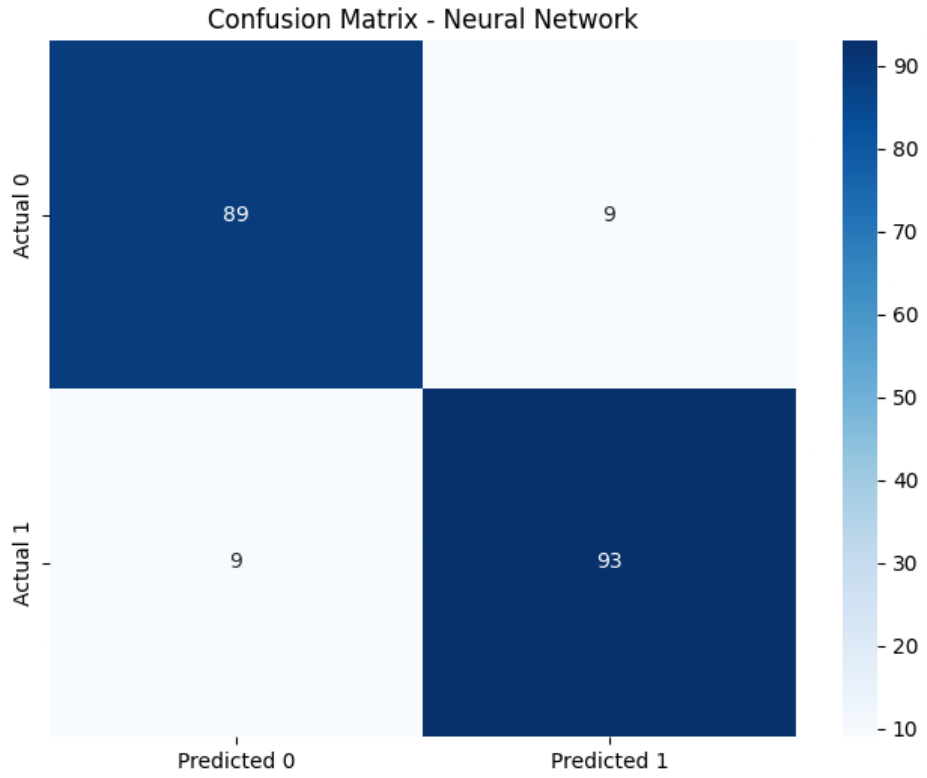
3) Model Neural Network

The Neural Network model also showed excellent performance with an accuracy rate **of 92%,** supporting the finding that deep learning methods are suitable for dropout predictions in educational data (Perdana Putra & Utami, 2024; Wan Yaacob et al., 2020).

Table 4. Neural Network Model Results

Metric	Value
Accuracy	0.92
Precision	0.92
Recall	0.91
F1-Score	0.91

Source: Processed by Authors (2023)



Gambar 5. Confusion matrix model Neural Network
Source: Processed by Authors (2023)

2. Model Performance Comparison

The results of the evaluation show that Random Forest consistently excels over Gradient Boosting and Neural Network in the context of dropout predictions based on demographic and academic data, as has been reported in several recent studies (Putra & Utami, 2024; Wan Yaacob et al., 2020; Sinaga, 2020).

A detailed comparison of model performance is presented in the following table:

Table 5. Comparison Results to Three Models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.93	0.93	0.92	0.92
Gradient Boosting	0.90	0.89	0.91	0.90
Neural Network	0.92	0.92	0.91	0.91

Source: Processed by Authors (2023)

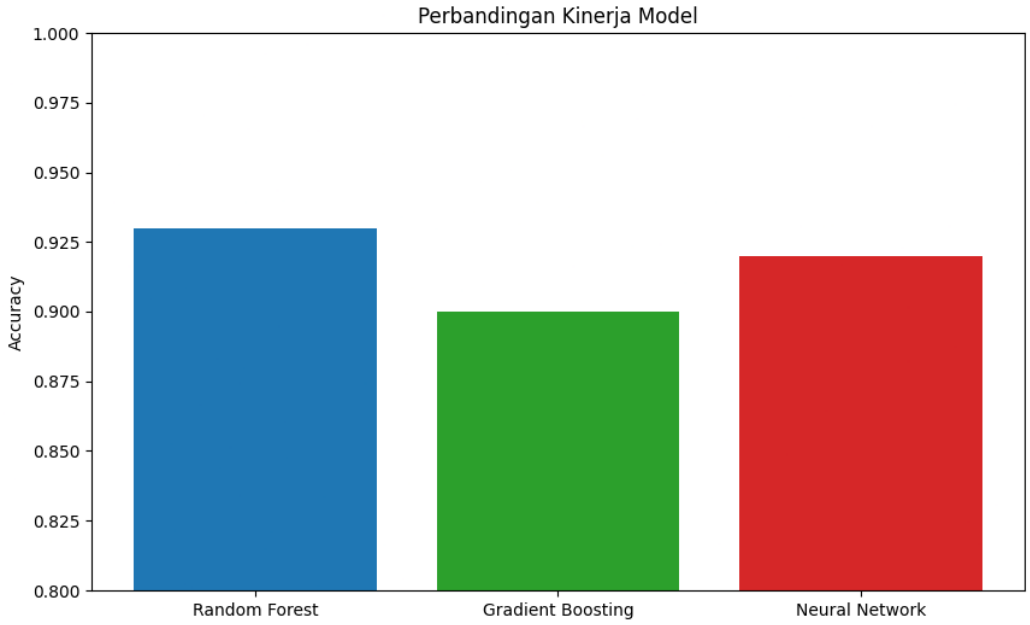


Figure 6. Comparison of the performance of three models of predicting the risk of dropping out of school

Source: Processed by Authors (2023)

3. The Importance of Predictive Variables

The feature importance analysis in Random Forest showed that the most influential variables were socioeconomic status, parental involvement, and academic performance. These findings are supported by many previous studies that also highlight the role of these factors as strong predictors of school dropout risk (Abdullah & Muhid, 2021; Fenech, 2024; Li, 2020; Agustina, 2019).

1) Socioeconomic status
2) Parental involvement
3) Performa akademik (Academic performance)

Table 6. Feature importance of the Random Forest model

Feature	Importance
distance_to_school	0.32
academic_performance	0.25
socioeconomic_status	0.20
parental_involvement	0.12
school_resources	0.11

Source: Processed by Authors (2023)

a. distance_to_school feature has the greatest influence on predicting the risk of dropping out of school.
b. **academic_performance** and **socioeconomic_status** also make significant contributions.

c. Other features such as **parental_involvement** and **school_resources** have a smaller but still important role.

CONCLUSION

This research successfully developed and compared machine learning models Random Forest, Gradient Boosting, and Neural Network—to predict school dropout risk. The Random Forest model demonstrated the highest accuracy (93%), outperforming the other models. Key factors influencing dropout risk included socioeconomic status, academic performance, and parental involvement. The findings highlight the potential of machine learning as an early warning system for schools, enabling targeted interventions to improve student retention. Future research could enhance predictive power by incorporating psychological variables and longitudinal data, as well as exploring additional algorithms for further optimization. Practical implementation of these models in schools could significantly reduce dropout rates and support educational equity.

REFERENCES

Abdullah, A. W., & Muhid, A. (2021). Social support, academic satisfaction, and student drop out tendency. Psikoislamika: Jurnal Psikologi dan Psikologi Islam, 18(1), 174–187. https://doi.org/10.18860/psi.v18i1.11546

Agustina, A. L. Y. (2019). Designing to improve student completion rates: Preliminary qualitative investigation into causes of drop-out. Desain Gakuronko, 15, 50–60.

Brandt, B. G. (2020). Finding their voices: An exploration of experiences that contribute to student drop-out in West Virginia higher education institutions (Doctoral dissertation, Edgewood College). ProQuest Dissertations Publishing.

Campbell, K. (2024). A suburban district's approach to addressing the student drop-out rate (Doctoral dissertation, Lindenwood University).

Cele, S. C., Diale, B. M., & Khumalo, S. (2025). Black African students' social and academic identities in South African universities vis-à-vis student drop out. Journal of Student Affairs in Africa, 13(1), 1–18. https://doi.org/10.24085/jsaa.v13i1.425

Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning. *Education and Information Technologies*, *28*(9), 12039–12057.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*(4), 498–506.

Do Nascimento, R. L. S., Fagundes, R. A. de A., & de Souza, R. M. C. R. (2022). Statistical learning for predicting school dropout in elementary education: A comparative study. *Annals of Data Science*, *9*(4), 801–828.

Fenech, F. (2024). Well-being of students enrolled in the field of education: A predictor for student drop-out? International Journal of Educational Research, 116, 102138. https://doi.org/10.1016/j.ijer.2024.102138

Guntara, M., & Suprawoto, T. (2024). Drop out student clusterization using the k-medoids algorithm. Jurnal Teknologi dan Sistem Komputer (JTKaSK), 12(1), 61–66. https://doi.org/10.33364/jtksk.v12i1.1081

Kim, S.-H., & Cho, S.-H. (2024). A Comparative Study of Prediction Models for College Student Dropout Risk Using Machine Learning: Focusing on the case of N university. *Journal of The Korean Society of Integrative Medicine*, *12*(2), 155–166.

Li, W. (2020). Factors behind rural student drop-out rates in North China: A qualitative study.

Cambridge Journal of China Studies, 14(4), 37–49.

Putra, M. R. P., & Utami, E. (2024). Comparative analysis of hybrid model performance using stacking and blending techniques for student drop-out prediction in MOOC. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 8(3), 346–354. https://doi.org/10.29207/resti.v8i3.5760

Sinaga, T. M. (2020). Aplikasi pengelompokan mahasiswa potensial drop out pada STMIK TIME. Jurnal TIMES, 9(1), 33–39. http://ejournal.stmik-time.ac.id

Vasconcelos, N., Júnior, M. C., Almeida, T., & da Silva, V. M. (2019). Comparative Analysis of Data Mining Algorithms Applied to the Context of School Dropout. *FedCSIS (Communication Papers)*, 3–10.

Villar, A., & de Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1), 2.

Wan Yaacob, W. F., Mohd Sobri, N., Md Nasir, S. A., Norshahidi, N. D., & Wan Husin, W. Z. (2020). Predicting student drop-out in higher institution using data mining techniques. Journal of Physics: Conference Series, 1496, 012005. https://doi.org/10.1088/1742-6596/1496/1/012005

Zawada, J. (2024). Student dropout and feelings of belonging and mattering in UK undergraduate allied health students. Journal of Student Success, 9(1), 15–28.

Agustian, F. H. (2019). Application of case-based reasoning for student recommendations drop out. Jurnal Teknologi dan Sistem Komputer, 7(2), 165–172.

Yohena, A. L. (2021). Drop-out teachers: Student composition and teacher mobility and attrition in lower secondary schools [Master's thesis, University of Oslo]. DUO Digital Archive. https://www.duo.uio.no/handle/10852/84763