Journal of Student Research Exploration

https://shmpublisher.com/index.php/josre

p-ISSN xxxx-xxxx | e-ISSN xxxx-xxxx

The impact of SMOTE and hyperparameter tuning on Random Forest for predicting student attrition

Nayla Dzilkamala[1], Isna Nadia Nuraini[2], Farida Maretia Niswah[3], Much Aziz Muslim[4]

[1,2,3,4] Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received May 15, 2026
Revised May 15, 2026
Accepted June 9, 2026

Keywords:

Student attrition
Random forest
SMOTE
Hyperparameter tuning
Machine learning

ABSTRACT

Student attrition remains a major challenge in higher education, requiring early intervention for at-risk students. This study examines the effect of Synthetic Minority Oversampling Technique (SMOTE) and hyperparameter tuning on Random Forest performance for predicting student attrition. Using 4,424 student records from the UCI Predict Students' Dropout and Academic Success dataset, the target variable was converted into binary classification (Dropout vs. Non-Dropout). Four Random Forest models were evaluated: RF-Baseline, RF-SMOTE, RF-Tuned, and RF-SMOTE-Tuned. Data were split into 80% training and 20% testing sets, while Grid Search with 5-fold Stratified Cross-Validation was applied for optimization. Performance was measured using accuracy, precision, recall, F1-score, and AUC. The RF-SMOTE-Tuned model achieved the best results with 0.8825 accuracy and 0.9314 AUC. Results show that SMOTE improved minority-class detection, while hyperparameter tuning increased model stability. Feature importance analysis identified approved curricular units, semester grades, and tuition fee status as the strongest predictors of student attrition.

1. Introduction

Student attrition has become a critical issue in higher education because it affects students' academic continuity, institutional performance, resource allocation, and long-term educational quality. Matz et al. emphasized that "Student attrition poses a major challenge to academic institutions, funding bodies and students" [1], [2]. Similarly, Deleña et al. highlighted that "Student dropout remains a persistent challenge in higher education" [3], [4].

[1] ***Corresponding Author:***
Nayla Dzilkamala,
Department of Computer Science,
State University of Semarang,
Sekaran, Gunungpati, Semarang City, Central Java 50229, Indonesia.
Email: nayladzil20@students.unnes.ac.id
DOI: https://doi.org/10.52465/josre.v4i2.7

These findings indicate that student attrition is not only an individual academic problem but also an institutional concern that requires systematic monitoring and early intervention.

The development of machine learning has provided new opportunities for higher education institutions to identify students who are at risk of dropping out. Vaarma and Li reported that "Learning management system data is powerful in degree dropout prediction" [5], [6]. In line with this, Dass et al. explained that "This paper presents a model to predict the dropout of students from a MOOC course" [7]. These studies show that educational data, learning behavior, and machine learning techniques can be used to support early identification of students who may discontinue their studies.

Early prediction is particularly important because it allows institutions to provide support before students leave their study programs. Lee and Chung argued that "A dropout early warning system enables schools to preemptively identify students who are at risk of dropping out" [8]. Kok et al. demonstrated that "This study leverages big data and machine learning to identify key parameters influencing student dropout" [9]. Therefore, student data such as academic performance, demographic background, socioeconomic status, and enrollment information can be utilized to develop predictive models that support student retention.

Accurate student attrition prediction can help universities design more effective intervention strategies. Opazo et al. defined dropout as "the abandonment of a high education program before obtaining the degree without reincorporation" [10]. Furthermore, Umendu et al. observed that "This study applies machine learning techniques to predict student non-continuation and attrition" [11]. Marcolino et al. underlined that "Student attrition and academic failure remain pervasive challenges in education" [12]. These findings strengthen the argument that predictive analytics can support higher education institutions in reducing dropout rates and improving student success.

Among various machine learning algorithms, Random Forest is widely used because it can handle complex relationships among variables and reduce the limitations of a single decision tree. Schonlau and Zou explained that "ensemble learning algorithms like random forests are well suited for medium to large datasets" [13], [14]. This ensemble structure makes Random Forest suitable for student attrition prediction because student datasets usually contain heterogeneous attributes, including academic, demographic, and socioeconomic variables. Flores et al. found that "the predictive model based on Random Forest is the one that presents the highest accuracy and robustness" [15]. However, the performance of Random Forest can still be influenced by data imbalance and hyperparameter settings.

Class imbalance is a common problem in student attrition datasets because the number of non-dropout students is often higher than the number of dropout students. Li et al. explained that "SMOTE is one of the most well-established oversampling algorithms" [16]. If class imbalance is not properly handled, the model may become biased toward the majority class and fail to identify students who are actually at risk of dropping out. To address this issue, SMOTE is commonly applied to generate synthetic samples for the minority class and improve the model's ability to detect dropout cases.

In addition to class imbalance, hyperparameter selection also plays an important role in improving Random Forest performance. Bergstra and Bengio noted that “Grid search and manual search are the most widely used strategies for hyper-parameter optimization” [17]. This means that using default Random Forest parameters may not always produce optimal results for a specific dataset. Therefore, hyperparameter tuning is necessary to find the most suitable model configuration and improve predictive performance.

Based on these previous studies, this research investigates the impact of SMOTE and hyperparameter tuning on Random Forest performance for predicting student attrition. This study evaluates four experimental configurations: Random Forest baseline, Random Forest with SMOTE, tuned Random Forest, and Random Forest with both SMOTE and hyperparameter tuning. This design is aligned with the study framework that isolates the individual and combined effects of SMOTE and hyperparameter tuning on student attrition prediction. The main contribution of this research is to determine whether class balancing, hyperparameter optimization, or the combination of both provides the most effective improvement in predicting student attrition.

2. Method

This study employs a quantitative computational experiment approach to evaluate the impact of SMOTE and hyperparameter tuning on Random Forest classification performance for predicting student attrition. The research pipeline consists of five sequential stages: data collection, data preprocessing, class imbalance handling, model training across four experimental configurations, and model evaluation. The overall research workflow is illustrated in Figure 1.

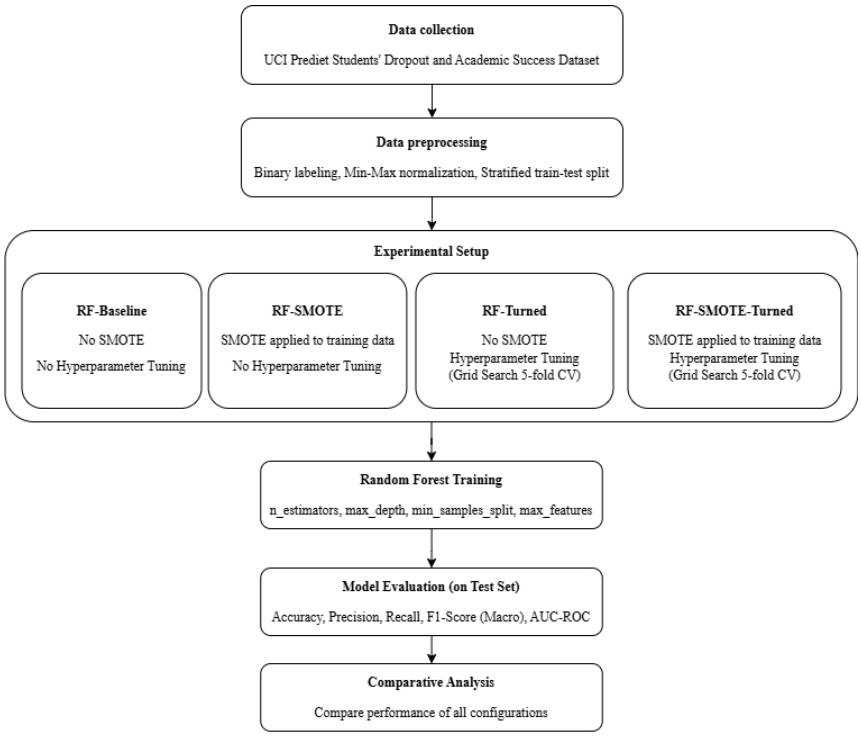

Figure 1. Research workflow of student attrition prediction using Random Forest, SMOTE, and hyperparameter tuning

2.1. Dataset

The dataset used in this study is the Predict Students' Dropout and Academic Success dataset publicly available through the UCI Machine Learning Repository [18]. The dataset comprises 4,424 student records with 36 attributes spanning three categories: demographic attributes including age at enrollment, gender, and nationality; socioeconomic attributes including parental educational background, scholarship holder status, and tuition fee payment status; and academic performance attributes including the number of curricular units credited, enrolled, evaluated, and approved in the first and second semesters along with the corresponding grade averages. The original target variable contains three classes: Dropout, Enrolled, and Graduate. In this study, the target is transformed into a binary classification task in which Dropout serves as the positive class and Non-Dropout, combining Graduate and Enrolled instances, serves as the negative class, focusing the prediction objective on identifying students at risk of leaving the institution. The resulting class distribution reflects an imbalance condition with 1,421 Dropout records (32.1%) and 3,003 Non-Dropout records (67.9%).

2.2. Data Preprocessing

Prior to model training, preprocessing steps are applied to ensure data quality and pipeline compatibility. The dataset contains no missing values. Min-Max Normalization is applied to all numerical features to scale attribute values into the range [0, 1] as expressed in Equation (1), ensuring that features with larger numerical ranges do not disproportionately influence the learning process [19].

$$\mathrm{X_{norm}} = \frac{\mathrm{X} - \mathrm{X_{min}}}{\mathrm{X_{max}} - \mathrm{X_{min}}} \quad (1)$$

The dataset is subsequently partitioned into 80% training data comprising 3,539 records and 20% testing data comprising 885 records using stratified random splitting to preserve the original class proportion in both partitions.

2.3. Handling Class Imbalance with SMOTE

The dataset exhibits a Dropout-to-Non-Dropout ratio of approximately 1:2.1. Training a classifier on this distribution without correction produces models biased toward the majority class, yielding high overall accuracy while underperforming on the minority dropout class [20], [21]. SMOTE [22] is applied exclusively to the training partition after the train-test split to prevent data leakage, as applying it prior to splitting would introduce synthetic samples into the test set and produce artificially optimistic evaluation results [23]. SMOTE generates each synthetic sample by selecting a minority-class instance, identifying its five nearest neighbors within the minority class, and interpolating a new instance along the connecting segment according to Equation (2), where $x_i$denotes the selected instance, $\hat{x}$denotes a randomly chosen neighbor, and $\lambda$is a random scalar in the range [0, 1].

$$x_{synthetic} = x_i + \lambda \cdot (\hat{x} - x_i) \quad (2)$$

This process is repeated until the training set reaches a balanced class distribution of 1:1.

2.4. Experimental Configurations

Four configurations of Random Forest are evaluated to isolate and quantify the individual and combined effects of SMOTE and hyperparameter tuning, as presented in Table 1. All configurations are trained on the same data partition and evaluated on the same test set to ensure a fair and consistent comparison.

Table 1. Experimental configurations of Random Forest

Configuration	SMOTE	Hyperparameter Tuning
RF-Baseline	No	No
RF-SMOTE	Yes	No
RF-Tuned	No	Yes
RF-SMOTE-Tuned	Yes	Yes

For RF-Baseline and RF-SMOTE, Random Forest is trained using scikit-learn default parameters. For RF-Tuned and RF-SMOTE-Tuned, hyperparameter optimization is performed using Grid Search with 5-fold Stratified Cross-Validation, with the configuration yielding the highest macro F1-score selected as the final model [19]. The hyperparameter search space is defined in Table 2.

Table 2. Hyperparameter search space for grid search

Hyperparameter	Values	Description
n_estimators	100, 200, 300	Number of trees in the ensemble
max_depth	None, 10, 20, 30	Maximum depth per tree
min_samples_split	2, 5, 10	Minimum samples to split a node
max_features	sqrt, log2	Features considered per split

2.5. Evaluation Metrics

All four configurations are evaluated on the original imbalanced test set to reflect real-world prediction conditions [24], [25]. Macro-averaged F1-Score serves as the primary metric as it weights each class equally regardless of support, making it more informative than overall accuracy under class imbalance. Accuracy, Precision, Recall, F1-Score, and AUC-ROC are computed according to Equations (3) through (6).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \tag{6}$$

All computational experiments are implemented in Python 3.10 using scikit-learn, imbalanced-learn, pandas, matplotlib, and seaborn, executed in Google Colaboratory to ensure reproducibility.

3. Results and Discussion

3.1. Experimental Results

This study evaluates four Random Forest configurations to analyze the impact of SMOTE and hyperparameter tuning on student attrition prediction performance. The evaluated configurations consist of RF-Baseline, RF-SMOTE, RF-Tuned, and RF-SMOTE-Tuned. All models were evaluated using the same imbalanced test set to ensure fair comparison under realistic conditions. The overall experimental results are presented in Table 3.

Table 3. Performance comparison of Random Forest configurations

Model	Accuracy	Precision	Recall	F1-Score	AUC
RF-Baseline	0.8836	0.8918	0.7254	0.8000	0.9292
RF-SMOTE	0.8814	0.8456	0.7711	0.8066	0.9278
RF-Tuned	0.8847	0.8824	0.7394	0.8046	0.9314
RF-SMOTE-Tuned	0.8825	0.8516	0.7676	0.8074	0.9314

The results indicate that all Random Forest configurations achieved strong predictive performance, with AUC values exceeding 0.92. Among all configurations, RF-SMOTE-Tuned achieved the highest F1-score of 0.8074, indicating the best balance between Precision and Recall for imbalanced student attrition prediction.

Although RF-Tuned achieved the highest Accuracy value, Accuracy alone is insufficient for evaluating imbalanced classification problems because it tends to favor the majority class. Therefore, Macro F1-score serves as the primary evaluation metric in this study as it equally considers minority and majority classes. The comparison of F1-score across all experimental configurations is illustrated in Figure 2.

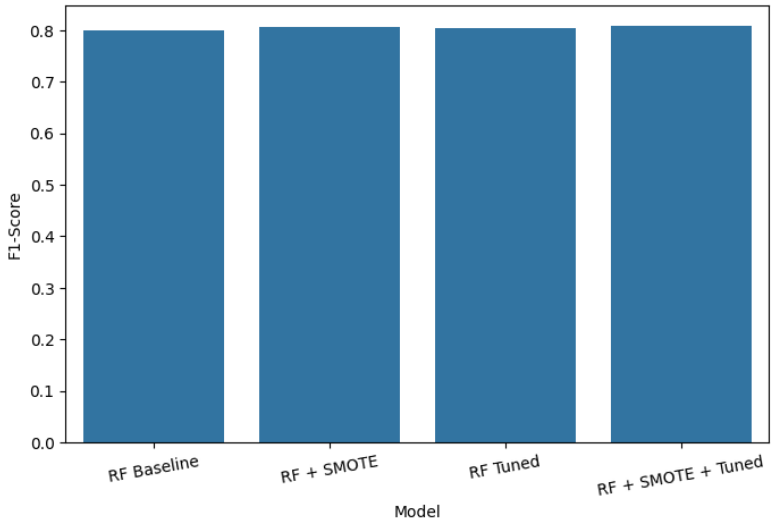

Figure 2. F1-Score comparison across Random Forest configurations

The visualization in Figure 2 demonstrates that the incorporation of SMOTE and hyperparameter tuning consistently improves classification balance. RF-SMOTE-Tuned produced the highest overall F1-score, confirming that combining data balancing and parameter optimization yields the most robust predictive performance. The implementation of SMOTE substantially improved Recall performance. Recall increased from 0.7254 in RF-Baseline to 0.7711 in RF-SMOTE, indicating that more at-risk students were successfully identified after balancing the minority class distribution. This finding confirms that class imbalance negatively affects dropout detection capability and that SMOTE effectively reduces this limitation. Hyperparameter tuning also improved model performance. RF-Tuned achieved higher F1-score and AUC values compared to RF-Baseline, indicating that parameter optimization enhances model generalization and classification stability.

3.2. Cross Validation Analysis

To further evaluate model robustness and generalization capability, 5-fold Stratified Cross Validation was applied to the best-performing configuration, RF-SMOTE-Tuned. The cross-validation results are presented in Table 4.

Table 4. 5-Fold cross validation results

Fold	F1-Score
Fold 1	0.8392
Fold 2	0.8389
Fold 3	0.8537
Fold 4	0.8578
Fold 5	0.8570
Mean	0.8493

The cross-validation results demonstrate stable model performance across different data partitions. The relatively small variation between folds indicates that the proposed model generalizes well and does not suffer from significant overfitting. The ROC curve of the best-performing model is illustrated in Figure 3.

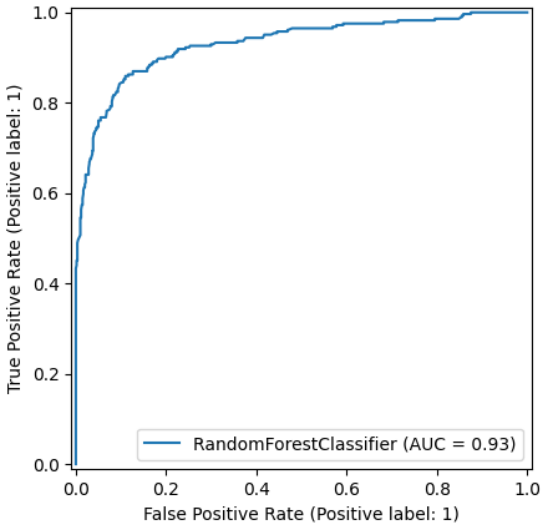

Figure 3. ROC curve of RF-SMOTE-Tuned model

Figure 3 shows that the ROC curve approaches the upper-left corner of the graph, indicating strong classification capability. The obtained AUC value above 0.93 confirms that the model effectively distinguishes dropout students from non-dropout students.

3.3. Confusion Matrix and False Negative Analysis

Confusion matrix analysis was conducted to further investigate model classification behavior, particularly in identifying dropout students. In student attrition prediction, False Negative (FN) cases are highly critical because they represent students who are actually at risk of dropping out but are incorrectly classified as non-dropout students. The confusion matrices for all Random Forest configurations are presented in Figure 4.

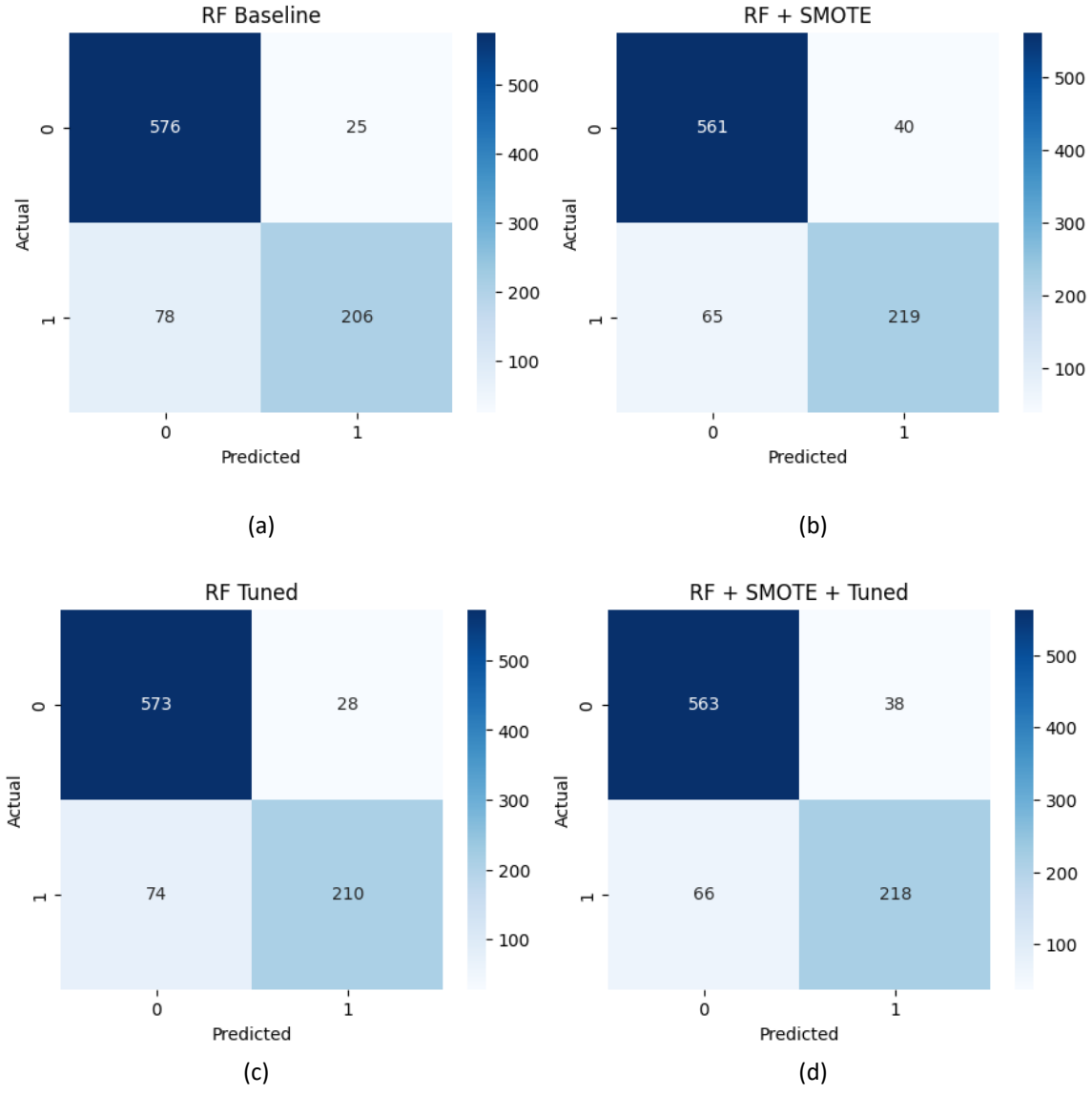

Figure 4. Confusion matrices for student attrition prediction using (a) the RF-Baseline model, (b) the RF-SMOTE model, (c) the RF-Tuned model, and (d) the RF-SMOTE-Tuned model

The RF-Baseline model produced a larger number of False Negatives compared to RF-SMOTE and RF-SMOTE-Tuned. After applying SMOTE, the number of missed dropout

students decreased significantly, indicating improved sensitivity toward minority-class detection. The comparison of False Negative counts between RF-Baseline and RF-SMOTE is illustrated in Figure 5.

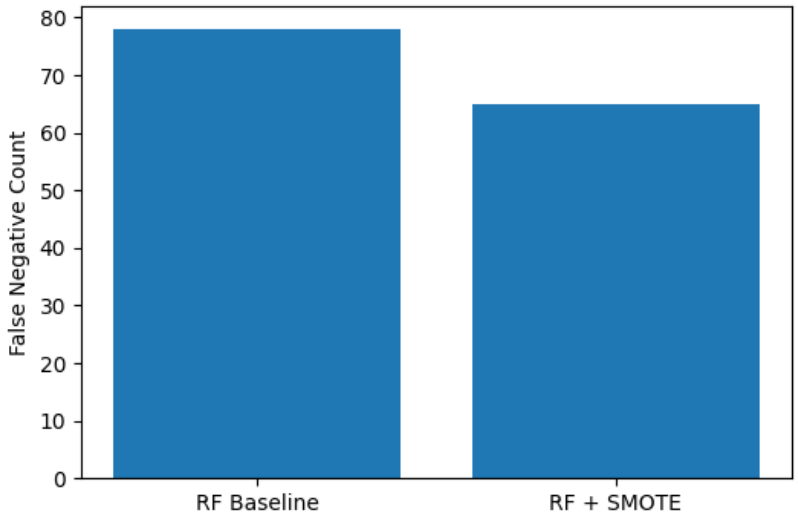

Figure 5. False negative comparison between RF-Baseline and RF-SMOTE

The reduction in False Negative cases demonstrates that balancing the training data enables Random Forest to better recognize complex dropout patterns that are underrepresented in the original dataset. Further analysis revealed that many False Negative students exhibited borderline academic characteristics rather than extremely poor academic performance. Several students still maintained moderate semester grades and partially completed curricular units, making them difficult to distinguish from non-dropout students using simpler classification patterns. From an educational perspective, this finding is highly important because borderline-risk students are frequently overlooked in institutional early warning systems. Identifying such students earlier may enable universities to provide timely interventions such as academic counseling, mentoring, or financial assistance before dropout occurs.

3.4. Feature Importance Analysis

Feature importance analysis was conducted using the RF-SMOTE-Tuned model to identify the variables that contribute most significantly to student attrition prediction. The ranking of the most influential features is illustrated in Figure 6.

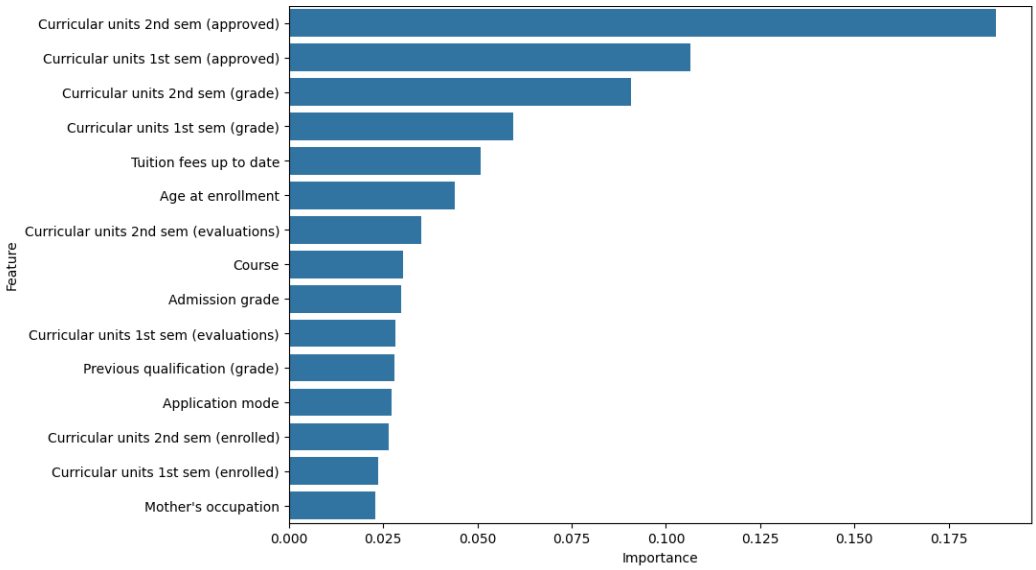

Figure 6. Top 15 Feature importance in student attrition prediction

The feature importance results indicate that academic performance variables dominate the prediction process. The most influential attributes include:

1. Curricular units 2nd semester (approved)
2. Curricular units 1st semester (approved)
3. Curricular units 2nd semester (grade)
4. Tuition fees up to date
5. Admission grade
6. Age at enrollment

The dominance of curricular unit approval and semester grade variables indicates that prior academic performance is one of the most consistent predictors of student retention, where students with stronger academic achievement are less likely to discontinue their studies [1].

Financial-related variables also exhibit considerable importance. The “Tuition fees up to date” feature consistently appears among the top predictors, suggesting that financial difficulties significantly influence dropout risk. Students experiencing delayed tuition payment may encounter economic pressure that negatively affects both academic engagement and educational continuity.

Additionally, admission grade and age at enrollment contribute meaningfully to prediction performance. Lower admission grades may indicate weaker academic preparedness prior to entering higher education, while older students may face additional occupational or socioeconomic responsibilities that increase dropout vulnerability. The correlation heatmap among variables is presented in Figure 7.

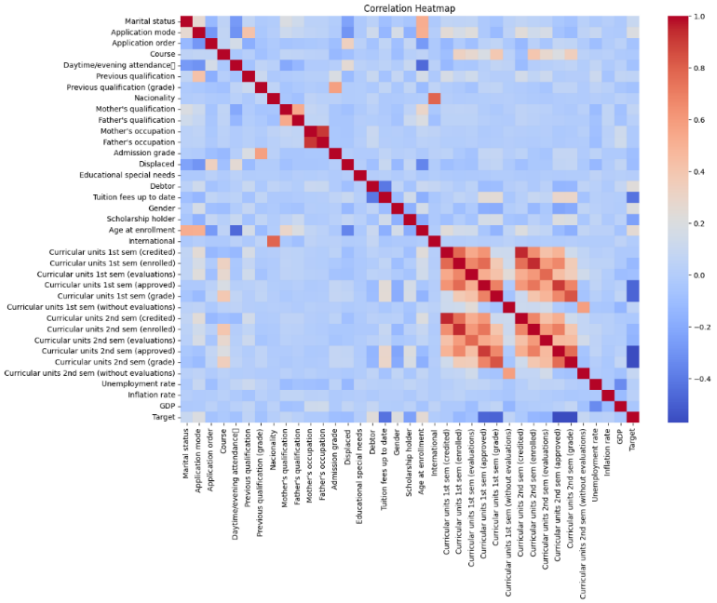

Figure 7. Correlation heatmap of student attributes

Figure 7 shows that academic performance attributes exhibit stronger correlations with attrition-related variables compared to demographic attributes. This finding further supports the conclusion that academic engagement and performance play dominant roles in determining student persistence.

Overall, the feature importance analysis provides meaningful educational insights beyond technical machine learning evaluation. Universities may utilize these findings to design targeted intervention strategies focused on students experiencing academic difficulties, financial instability, or early signs of disengagement.

3.5. Discussion of Model Behavior

The performance differences among the four experimental configurations can be explained by the effects of class balancing and parameter optimization on the Random Forest learning process. The RF-Baseline model demonstrated strong overall performance; however, its Recall value remained relatively lower due to the imbalanced class distribution. In imbalanced educational datasets, machine learning models tend to become biased toward the majority class, causing dropout students to be underdetected.

The implementation of SMOTE improved minority-class representation by generating synthetic dropout samples within the training data. As a result, the RF-SMOTE configuration achieved higher Recall and F1-score values, indicating improved capability in identifying at-risk students. This finding demonstrates that balanced training data help Random Forest learn more representative dropout patterns and reduce the number of False Negative cases.

Hyperparameter tuning further improved model stability and generalization capability. By optimizing parameters such as the number of trees, maximum tree depth, and feature selection strategy, the Random Forest model was able to produce more robust predictions across unseen student records. The RF-SMOTE-Tuned configuration achieved the highest

overall F1-score, confirming that combining data balancing and parameter optimization provides the best classification performance for student attrition prediction.

The feature importance analysis also revealed that academic performance variables dominated the prediction process. Prior academic performance has been identified as one of the most consistent predictors of student retention, where students with stronger academic achievement are less likely to discontinue their studies. This finding indicates that academic engagement remains a critical factor in determining student persistence in higher education environments.

3.6. Implications for Early Warning System

The results of this study demonstrate that machine learning models can effectively support early warning systems for predicting student attrition risk. The superior performance of the RF-SMOTE-Tuned model, particularly in improving Recall and reducing False Negative cases, makes it more suitable for institutional implementation. Minimizing False Negatives is highly important because students incorrectly classified as non-dropout students may fail to receive timely academic intervention despite being at risk.

The findings further suggest that universities should prioritize continuous monitoring of students' academic performance and financial conditions. Variables such as approved curricular units, semester grades, and tuition fee status consistently contributed to dropout prediction, indicating that both academic disengagement and financial difficulties are strongly associated with attrition risk.

By accurately identifying vulnerable students during the early stages of study, educational institutions may provide targeted interventions such as academic counseling, mentoring programs, financial assistance, and personalized student support services. Consequently, predictive analytics models may contribute to improving student retention, reducing dropout rates, and supporting data-driven decision-making in higher education institutions.

4. Conclusion

This study investigated the impact of SMOTE and hyperparameter tuning on Random Forest performance for predicting student attrition using the UCI Predict Students' Dropout and Academic Success dataset. Four experimental configurations were evaluated, namely RF-Baseline, RF-SMOTE, RF-Tuned, and RF-SMOTE-Tuned. The experimental results demonstrate that the integration of SMOTE and hyperparameter tuning improves classification performance, particularly in identifying dropout students within imbalanced educational datasets. Among all configurations, RF-SMOTE-Tuned achieved the best overall performance with the highest F1-score of 0.8074 and AUC value above 0.93, indicating strong classification capability and balanced minority-class detection.

The findings further reveal that SMOTE significantly improves Recall performance and reduces False Negative cases, which is critical in educational early warning systems because undetected at-risk students may fail to receive timely intervention. In addition, feature importance analysis showed that academic performance variables, particularly approved

curricular units and semester grades, are the most influential predictors of student attrition. Overall, this study confirms that combining data balancing techniques and hyperparameter optimization enhances Random Forest effectiveness for student attrition prediction. The proposed approach may support universities in developing data-driven early warning systems to improve student retention and reduce dropout rates.

Despite these findings, this study has several limitations, including the use of a single public dataset and static academic attributes. Future research may incorporate temporal behavioral data, compare additional ensemble learning methods, or explore explainable artificial intelligence techniques such as SHAP for deeper educational interpretation.

REFERENCES

[1] S. C. Matz, C. S. Bukow, H. Peters, C. Deacons, and C. Stachl, "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," *Sci. Rep.*, vol. 13, no. 1, pp. 1–16, 2023, doi: 10.1038/s41598-023-32484-w.

[2] B. Duro, A. Gomes, F. B. Correia, A. R. Borges, and J. Bernardino, "Machine Learning and Deep Learning for Dropout Prediction in Higher Education: A Review," *Computers*, vol. 15, no. 3, pp. 1–26, 2026, doi: 10.3390/computers15030164.

[3] R. D. Deleña *et al.*, "Predicting student retention: A comparative study of machine learning approach utilizing sociodemographic and academic factors," *Syst. Soft Comput.*, vol. 7, no. July 2024, 2025, doi: 10.1016/j.sasc.2025.200352.

[4] A. Bettahi, F. Z. Belouadha, and H. Harroud, "A Modular and Explainable Machine Learning Pipeline for Student Dropout Prediction in Higher Education," *Algorithms*, vol. 18, no. 10, pp. 1–31, 2025, doi: 10.3390/a18100662.

[5] M. Vaarma and H. Li, "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, no. September 2023, p. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.

[6] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, and R. M. G. García-Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Front. Educ.*, vol. 8, no. October, pp. 1–14, 2023, doi: 10.3389/feduc.2023.1244686.

[7] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced mooc course using random forest model," *Inf.*, vol. 12, no. 11, 2021, doi: 10.3390/info12110476.

[8] S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Appl. Sci.*, vol. 9, no. 15, 2019, doi: 10.3390/app9153093.

[9] C. L. Kok, C. K. Ho, L. Chen, Y. Y. Koh, and B. Tian, "A Novel Predictive Modeling for Student Attrition Utilizing Machine Learning and Sustainable Big Data Analytics," *Appl. Sci.*, vol. 14, no. 21, 2024, doi: 10.3390/app14219633.

[10] D. Opazo, S. Moreno, E. Álvarez-Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," *Mathematics*, vol. 9, no. 20, pp. 1–27, 2021, doi: 10.3390/math9202599.

[11] E. C. Umendu, M. Ghanzanfar, A. Kans, and M. A. R. Ahad, "Enhancing Student Retention in Higher Education Institutions (HEIs): Machine Learning Approach," *Electron.*, vol. 15, no. 4, 2026, doi: 10.3390/electronics15040734.

[12] M. Rebelo Marcolino *et al.*, "Student dropout prediction through machine learning optimization: insights from moodle log data," *Sci. Rep.*, vol. 15, no. 1, pp. 1–16, 2025, doi: 10.1038/s41598-025-93918-1.

[13] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020, doi: 10.1177/1536867X20909688.

[14] U. Olive, M. J. Bosco, and N. M. Enan, "Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis," *J. Inf. Technol.*, vol. 5, no. 4, pp. 31–40, 2025, doi: 10.70619/vol5iss4pp31-40.

[15] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electron.*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030457.

[16] Y. Li, Y. Yang, P. Song, L. Duan, and R. Ren, "An improved SMOTE algorithm for enhanced imbalanced

data classification by expanding sample generation space," *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-09506-w.

[17] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012.

[18] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, no. 11, 2022, doi: 10.3390/data7110146.

[19] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, pp. 1–19, 2019, doi: 10.1002/widm.1301.

[20] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

[21] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[23] T. A. Marzuqi, E. Kristiani, and Marcel, "Prediksi Mahasiswa Drop-Out Di Universitas XYZ," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 6, pp. 1345–1350, 2024, doi: 10.25126/jtiik.2024118689.

[24] N. Mduma, K. Kalegele, and D. Machuve, "A survey of machine learning approaches and techniques for student dropout prediction," *Data Sci. J.*, vol. 18, no. 1, pp. 1–10, 2019, doi: 10.5334/dsj-2019-014.

[25] I. M. S. Bimantara, I. W. Supriana, and I. K. G. Suhartana, "Strategi optimalisasi hyperparameter model machine learning untuk prediksi putus studi dini mahasiswa," *JELIKU (Jurnal Elektron. Ilmu Komput. Udayana)*, vol. 14, no. 1, pp. 141–156, 2025, [Online]. Available: https://ojs.unud.ac.id/index.php/jlk/article/view/130088